Parteka, Aleksandra; Płatkowski, Piotr; Szymczak, Sabina; Wolszczak-Derlacz, Joanna

Working Paper

A dataset on knowledge creation and patenting by European Higher Education Institutions (KC-HEI)

GUT FME Working Paper Series A, No. 2/2024 (73)

Provided in Cooperation with:
Gdańsk University of Technology, Faculty of Management and Economics

Suggested Citation: Parteka, Aleksandra; Płatkowski, Piotr; Szymczak, Sabina; Wolszczak-Derlacz, Joanna (2024) : A dataset on knowledge creation and patenting by European Higher Education Institutions (KC-HEI), GUT FME Working Paper Series A, No. 2/2024 (73), Gdansk University of Technology, Faculty of Management and Economics, Gdansk

This Version is available at:
https://hdl.handle.net/10419/300341

GDAŃSK UNIVERSITY OF TECHNOLOGY

FACULTY OF MANAGEMENT AND ECONOMICS

A DATASET ON KNOWLEDGE CREATION AND PATENTING BY EUROPEAN HIGHER EDUCATION INSTITUTIONS (KC-HEI)

Aleksandra Parteka*, Piotr Płatkowski**

Sabina Szymczak*, Joanna Wolszczak-Derlacz******

GUT Faculty of Management and Economics

Working Paper Series A (Economics, Management, Statistics)

No 2/2024 (73)

June 2024

* Gdańsk University of Technology, Faculty of Management and Economics,
Narutowicza 11/12, 80-233 Gdańsk, Poland, aparteka@zie.pg.edu.pl (corresponding author)
** Gdańsk University of Technology, Faculty of Management and Economics,
Narutowicza 11/12, 80-233 Gdańsk, Poland, piotr.platkowski@pg.edu.pl
*** Gdańsk University of Technology, Faculty of Management and Economics,
Narutowicza 11/12, 80-233 Gdańsk, Poland, sabina.szymczak@pg.edu.pl
**** Gdańsk University of Technology, Faculty of Management and Economics,
Narutowicza 11/12, 80-233 Gdańsk, Poland, jwo@zie.pg.edu.pl

A dataset on knowledge creation and patenting by European Higher Education Institutions (KC-HEI)[1]

Aleksandra Parteka*, Piotr Płatkowski**
Sabina Szymczak***, Joanna Wolszczak-Derlacz****

June 26, 2024

Abstract

This paper describes the construction of a microlevel database on knowledge creation by higher education institutions (KC-HEI), accompanying the Global Knowledge Input-Output database (KIO, Davies et al., 2023). The database was created as part of Project Rethink GCS. KC-HEI links PATSTAT information on the patenting activity of 866 universities (HEIs) in 31 European countries over four decades (1980-2019), using citation records and patent quality indicators from OECD/STI Micro-data. KC-HEI makes possible analysis of the Institutions' innovation performance across 128 internationally comparable technological sectors and, separately, with respect to Artificial Intelligence (AI). We also develop a unique crosswalk between PATSTAT and ETER that combines KC-HEI with other institution-level datasets (such as ETER and RISIS) and allows us to build a parallel dataset covering 785 patenting and 2101 non-patenting universities in Europe between 2011 and 2019. We illustrate the potential of the KC-HEI database, providing key stylised facts on the role of universities in knowledge creation, while documenting extreme core-periphery patterns of university patenting in Europe and detecting several key university-level factors that reinforce this disparity.

JEL: O31, O33, I23

Keywords: Patents, Innovation, Knowledge, Higher Education Institutions, University

[1] This work was funded by the European Union under Horizon Europe grant 101061123 (Project Rethink-GSC). However, the views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. All errors are our own.
*Gdansk University of Technology. Narutowicza 11/12. Gdańsk. Poland. email: aparteka@zie.pg.edu.pl (corresponding author).
**Gdansk University of Technology. Narutowicza 11/12. Gdańsk. Poland. email: piotr.platkowski@pg.edu.pl
*** Gdansk University of Technology. Narutowicza 11/12. Gdańsk. Poland. email: sabina.szymczak@pg.edu.pl
**** Gdansk University of Technology. Narutowicza 11/12. Gdańsk. Poland. email: jwo@zie.pg.edu.pl

1. Introduction

This paper describes the creation of a new micro-level dataset designated KC-HEI, which accompanies the Knowledge Input Output database (KIO, Davies et al., 2023) and allows for a detailed analysis of the knowledge inputs (patents) provided by European higher education institutions (HEIs[2]) since 1980.

We focus on universities' patenting activity as a proxy for innovation and knowledge creation (Ayerst et al., 2023; Davies et al., 2023; Nagaoka et al., 2010), while taking account of all the caveats on the use of patents as an innovation indicator (Nagaoka et al., 2010). As to knowledge creators, patents reflect the innovation activity and new knowledge developed by firms (Caviggioli et al., 2023a; Lotti & Marin. 2013; Behrens & Trunschke, 2020; Chalioti et al., 2020; Aiello et al., 2022) and/or in universities/research institutions (Henderson et al., 1998; Cesaroni & Piccaluga. 2002; Coupé. 2003; Gurmu et al., 2010; Squicciarini et al., 2012; Whalley & Hicks. 2014; Duarte et al. 2020; Caviggioli et al., 2023a; Caviggioli et al., 2023b; Angori et al., 2023). We focus on the European university sector, trying to understand why the higher education sector plays only a minor role in market-oriented innovation within the EU. This issue relates closely to the current EU strategies for innovation, which are intended to strengthen universities as "drivers of the EU's global role and leadership" and acknowledge their fundamental role in furthering cooperation in research and innovation (EC, 2022a; EC, 2022b).

The evidence is cause for some concern. The European Patent Office (EPO, 2024)[3] reports that 69% of all patent applications in European countries were filed by large companies, 23% by SMEs and individual inventors, and 8% by universities and public research organisations.[4] Patenting is one of the ways through which universities realise their so-called "third mission" (Etzkowitz and Leydesdorff, 1997; Compagnucci and Spigarelli, 2020), after teaching and research, but in Europe not even 10% of patents are jointly filed by firms and universities. [5] Additionally, the international patenting position of European HEIs is weak and the debate on the "European university paradox"/"European paradox"[6]

[2] For simplicity, the terms Higher Education Institution and university are used interchangeably throughout. Formally, HEIs are defined by the European Tertiary Education Register (ETER; www.eter-project.com) as entities granting degrees at the tertiary level (ISCED levels 5 to 8). We keep ISCED 5 institutions (providing short-cycle tertiary education, such as German Fachhochschule, in the sample because they do report patents and often provide higher levels of education as well (ISCED 6 or 7).

[3] https://www.epo.org/en/about-us/statistics/patent-index-2023/statistics-and-indicators/applicants/categories [date of access: May 25, 2024].

[4] In the RISIS Patent database the university share is about 10%. The small share of HEIs among patent applicants is a global tendency: almost 88% of all Patent Cooperation Treaty (PCT) applications filed in 2022 came from the business sector (WIPO, 2023: 29).

[5] Own calculations based on a sample of HEIs from 17 European countries. (AT, BE, CH, CZ, DE, ES, FI, FR, HU, IE, IT, LT, LV, NL, PL, PT, UK) between 2011 and 2018. Primary data source: RISIS Patent database.

[6] This term was coined by the EC (1995) to describe the inability of European universities to translate top-level research into market-oriented innovations and competitive advantages. Later on, the paradox was questioned by authors who notede that Europe lags behind the US not only in knowledge commercialisation and university-industry cooperation, but also in top-quality scientific output gauged by other indicators, not just the number of scientific publications (among others, Cont. and Gaule, 2011; Dosi et al., 2006; Rodríguez-Navarro and Narin, 2018).

(EC, 1995; Conti and Gaule, 2011; Dosi et al., 2006; Rodríguez-Navarro and Narin, 2018) seems to be still open because European universities lag behind the global leaders in patenting. The global WIPO ranking of universities in 2022 was dominated by US, Korean, and Chinese institutions: among the world's top 50 patenting universities in 2020-2022[7] we find 18 universities based in the United States (led by the University of California) and 18 in China but only two in Europe (both in the UK)[8]. The top 10 in the Sciamago ranking[9] (2024) count eight Chinese and two US universities.

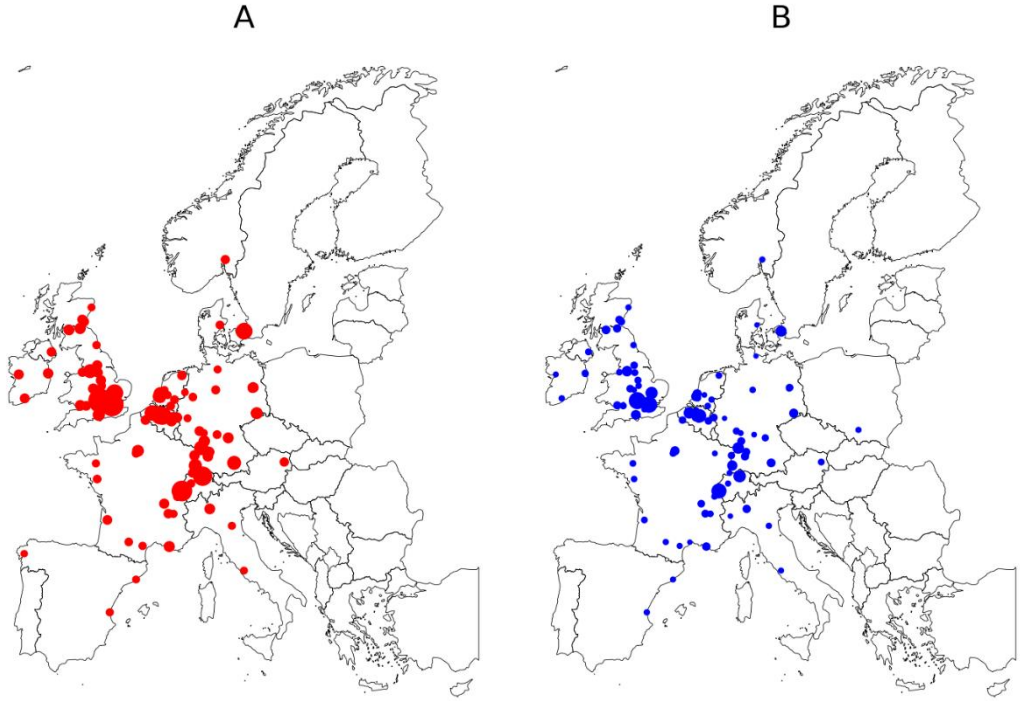


Figure 1. Top 100 patenting universities in Europe (1980-2019)

Source: own elaboration using KC-HEI
Note: The 100 universities with the most IP5 patent applications (Fig A.) and IP5 patents granted (Fig. B) in 1980-2019. Sample: 866 patenting HEIs in 31 European countries. IP5 patents allocated to HEIs using fractional apportionment by applicant share. Bubble size reflects the total number of patent applications/patents by university in 1980-2019.

Additionally, there is an extremely high geographical concentration of university patenting. KC-HEI data show that the top patenting universities are located in the economic heartland of Western Europe, mainly in the UK, Germany, France, Benelux and Switzerland (Figure 1). The extreme polarisation of university patenting in Europe is worrisome: 70% of all university patents come from five Western European countries, while over 70% of European HEIs show no patented knowledge creation at all. Fifty universities account for almost 27% of all university patent applications to the "big

[7] Top 50 university PCT applicants, 2020–2022. Source: WIPO Statistics Database, March 2023 (https://www3.wipo.int/ipstats/key-search/indicator).

[8] Namely: technology commercialisation and investment companies linked to the Imperial College of London (Imperial College Innovations Ltd) and Oxford University (Oxford University Innovation Limited).

[9] The Sciamago innovation ranking counts the number of patent applications by institutions and the citations that its research output receives from patents:
https://www.scimagoir.com/rankings.php?sector=Higher+educ.&ranking=Innovation&country=all [date of access: May 29, 2024].

five" patent offices.[10] Even within STEM institutions,[11] which should theoretically be better equipped to deliver market-oriented innovation in patentable technological fields, only 50% are active in patenting.

The regional distribution of HEIs' patenting activity is important because of the relationship between university-led R&D, regional inequalities and the convergence process: the quality of university research acts as a driver for regional per capita GDP growth (Agasisti and Bertoletti, 2022). The effect is likely to be two-way: the good economic performance of highly developed regions can enhance patenting by HEIs, which in turn may impact positively on local innovation potential (Caviggioli et al., 2023b; Bilbao-Osorio and Rodriguez-Pose; 2004).

To quantify the role of the university sector in innovation and knowledge flows, and to understand the determinants of the core-periphery pattern of university patenting in Europe, we need to develop granular data on patent records and the characteristics of university units. This is no trivial task, because patent-level records from sources such as PATSTAT do not provide coherent identification of universities among the patent assignees that can be merged with other datasets containing the characteristics of university units. A similar problem concerns research on firms, i.e. the difficult match between patent records and company-level data such as Amadeus or Orbis (Bremer, 2023).

Our KC-HEI dataset gives information on four decades of patenting activity (1980-2019) by nearly 900 universities in 31 European countries, which we then compare with a large sample of some 2100 non-patenting universities[12]. KC-HEI complements alternative entity-level data sources on patenting by universities such as RISIS Patent[13] and OrgReg[14], extending their coverage and increasing the degree of detail. RISIS Patent and OrgReg lack some crucial information on patent quality, whereas KC-HEI provides 30 alternative university-level indicators of the quality of the knowledge patented, based on the information on patent citations, breakthroughness, originality, etc.[15] We also provide parallel indicators based on universities' activity in alternative patent offices (IP5, EPO, USPTO), considering all patent applications and also only successful ones. What is more, KC-HEI permits analysis of the institutions' innovation performance in 128 internationally comparable technological

[10] Accounting for fractional apportionment. Numbers based on KC-HEI - see Section 3 for details.

[11] STEM (Science, Technology, Engineering, Mathematics) institutions are identified in line with ETER (Lepori, 2023: 119) as HEIs with positive shares of students and graduates in science and technology, i.e. in the following fields: 05 (Natural sciences, mathematics and statistics), 06 (Information and communication technologies) and 07 (Engineering, manufacturing and construction).

[12] The non-patenting sample can only be constructed for a shorter period, namely 2011-2019. Details in Section 2.4.

[13] Research Infrastructure for Science and Innovation Studies. RISIS Patent is managed by the Research Infrastructure for Science and Innovation Policy Studies (RISIS2.EU) project (https://www.risis2.eu/risis-datasets/). The data is provided upon acceptance – by dataset producers and the RISIS project review board – of a short project prepared by the applicant, who must register and agree on the terms governing its use.

[14] Register of Research and Higher Education Organizations (https://www.risis2.eu/orgreg-data/). All users with an active RISIS account have access to the OrgReg data.

[15] For details on all the indicators in KC-HEI see Appendix A.

sectors, and we further heighten the technological detail of analysis with separate university-level indicators of artificial intelligence (AI) patenting. To the best of our knowledge, this is the first attempt to provide a microlevel dataset that can quantify AI patenting by universities in such a large sample of institutions in different countries. So far AI patenting (or patenting in the so-called "4th Industrial Revolution" domain) has mainly been explored at the level of countries (Venturini, 2022; Parteka and Kordalska, 2023), regions (Balland and Boschma, 2021) and firms (Benassi et al., 2022; Czarnitzki et al., 2023; Yang, 2022; Igna and Venturini, 2023) but not explicitly higher education institutions. Lastly we contribute by providing a unique crosswalk between the codes (and names) of universities as patent assignees in PATSTAT and university identifiers in external datasets such as ETER. This tool allows us (and, hopefully, other researchers as well) to shed light on the characteristics of universities as knowledge creators in the European patenting system and formulate policy implications based on this highly granular data.

The rest of the paper is structured as follows. Section 2 describes the construction of the database in detail. Section 3 demonstrates the potential of KC-HEI, presenting a set of stylised facts on university patenting in Europe (1980-2019) and showing the core-periphery pattern of knowledge creation in this sector. Section 4 concludes. The Appendices set forth details on the content of KC-HEI (Appendix A), a detailed description of the key codes/procedures in its creation (Appendix B), and additional tables and figures (Appendix C).

2. KC-HEI dataset construction[16]

2.1 Data sources[17]

To quantify the creation of knowledge by universities, we use granular patent statistics from the Worldwide Statistical Patent Database (PATSTAT Global, Autumn 2022 edition), which gives bibliographical and legal event patent data from leading industrialised and developing countries. PATSTAT provides patent-level records allowing the quantification of innovation activity by firms (Caviggioli et al., 2023a; Lotti & Marin. 2013; Behrens & Trunschke. 2020; Chalioti et al., 2020; Aiello et al., 2022) and universities (among others: Squicciarini et al., 2012; Caviggioli et al., 2023a; Caviggioli et al., 2023b; Angori et al., 2023). We use the following data files from PATSTAT: Tls201_appln.csv. Tls206_person.csv. Tls207_pers_appln.csv. Tls209_appln_ipc.csv. Tls224_appln_cpc.csv.[18] We rely on the original PATSTAT data, not datasets derived from it, such as OECD REGPAT (in the university context used by, among others, Graf and Menter, 2022), because it offers much more granular information and facilitates the identification of university patents in a large sample of HEIs in various countries using the information on applicants' sector.[19] Google Patents, for its part, does not allow large-sample identification of universities among patent assignees, and its data is much less detailed than that of PATSTAT.

University identifiers and names come from the European Tertiary Education Register database (ETER)[20] (Lepori et al., 2023). ETER is hailed as the most comprehensive source of university-level data on such dimensions as institutional characteristics, geographical descriptors, details on education and research activities, financial records (expenditures and revenues), academic and non-academic personnel, and so on[21]. We also use ETER to create, as an integral part of the KC-HEI, a crosswalk

[16] A complete replication package accompanying this paper will be available at https://doi.org/10.34808/jzqd-zr04

[17] For details on provenance and availability of all the data, see the ReadmeFile (replication package).

[18] We thank Ronald Davies for his help in providing PATSTAT source files.

[19] OECD REGPAT, derived from PATSTAT, provides information on EPO and PCT patents by region. It has the advantage of free access and detailed data regionalisation (OECD, 2022). PATSTAT, on the other hand, provides information also on applications to other patent offices, including the major ones like JPO, KIPO and CNIPA. What is crucial for our study is identification of universities among the patent assignees. So far, this has been done with OECD REGPAT data through simple algorithms (Belvončíková, 2021 on selected European universities) or manual checks (Graf & Menter, 2022 on German universities). The latter method yields better matching quality but is time-consuming, so the samples are often limited to a single country. Our approach combines complex matching algorithms with manual checking to assure the most accurate correspondence and to limit the loss of relevant observations. Thanks to the use of the *psn_sector* variable from PATSTAT, which identifies universities among assignees in the first step, the entire process remains manageable even in a large sample.

[20] Data source: ETER project. Download date: September 18, 2023. Data has been provided by the European Tertiary Education Register (ETER) funded by the European Commission under the contracts EAC-2013-0308. EAC-2015-0280. 934533-2017 A08-CH and EAC-2021-0170.

[21] In particular, ETER provides: "Institutional descriptors. including legal status, institutional category, foundation year, etc.; Geographical descriptors, including the region of establishment, the city and the geographical coordinates of the institution; Educational activities: data on students and graduates by level of education, diploma, bachelor, master; ISCED5), field of education, gender, citizenship, mobility, age groups, full-time/part-time and the number of incoming and outgoing Erasmus students; Research activities: research-active institution, Ph.D. students and graduates (I.CED8), R&D expenditures; Expenditures, divided between personnel, non-personnel and capital, and revenues, divided between core budget, third-party funding and student fees funding; Personnel: academic personnel by gender, citizenship and field of education;

that relates university units in PATSTAT with external university-level datasets, using the *eter_id* identifier (details in Section 2.2).

Indicators of patent quality by application cohort (i.e. patents with the same filing year and technological class) are derived from the OECD Patent Quality Indicators database (version: August 2023) available at the OECD STI Micro-data Lab[22]. This dataset contains a series of indicators that capture the technological and economic value of EPO and USPTO patents (Squicciarini et al., 2013) and will allow us to accompany the basic indicator of patenting activity (number of patent applications) with such measures as the number of breakthrough patents in a university's portfolio and the number of patent citations.

Technological fields of patents and their Cooperative Patent Classification (codes CPC) are sourced from PATSTAT. We also use WIPO's PATENTSCOPE Artificial Intelligence Index[23] to identify university AI patents (i.e. those ascribable to artificial intelligence technologies). Finally, we use socio-economic data on the regions in which universities are located from the Eurostat Regional Statistics, NUTS 2 and NUTS3 level.[24]

2.2 Patent data selection

To construct KC-HEI, we use a subsample of 333,495 patent records reported in PATSTAT Global (Autumn 2022 edition) that can be ascribed to higher education units (i.e. applications with at least one university among the assignees[25] - details below) and with at least one applicant in a European country[26] since 1980. The final number of applications that serve to derive KC-HEI is smaller (107,501) because of choices concerning patent provenance and type. The choices concerning the final selection of countries and patent types are relatively easy: a considerably more complicated issue relates to the identification of university units in PATSTAT.

First, the final set of 31 countries (see Section 2.6) reflects the fact that we retain only those universities that can also be identified in ETER. This makes it possible to merge the PATSTAT records with university-level characteristics derived from external databases. In some countries – in

support and administrative personnel; research and teaching assistants; full professors by gender. A set of characterisation indicators concerning gender, citizenship, mobility, composition of personnel and HEI revenues; Information about demographic events in order to track institutions over time and observe development in the higher education sector." (Source: ht.ps://www.eter-project.com/overview-data/, assessed on October 25, 2023).

[22] https://www.oecd.org/sti/intellectual-property-statistics-and-analysis.htm. We use the following files:
202308_OECD_PATENT_QUALI.Y_EPO_INDIC.txt,
202308_OECD_PATENT_QUALITY.EPO_INDIC_COHORT.txt,
202308_OECD_PATENT_QUALITY_USPTO_INDIC.txt,
202308_OECD_PATENT_QUALITY_USPTO_INDIC_COHORT.txt.

[23] https://www.wipo.int/tech_trends/en/artificial_intelligence/patentscope.html [date of access: March 27, 2024]

[24] https://ec.europa.eu/eurostat/web/regions/database [date of access: March 27, 2024].

[25] Records in PATSTAT that have an assignee where PSN_Sector includes UNIVERSITY, i.e. "UNIVERSITY", "GOV NON-PROFIT UNIVERSITY", "UNIVERSITY HOSPITAL", "COMPANY UNIVERSITY". Later we consider the UNIVERSITY units only.

[26] We have selected patent applications with at least one assignee from the following countries: (variable *person_ctry_code* in TLS206 table in PATSTAT): AL AT BE BG CH CY CZ CS DE DD DT DL DK EE ES FI FR GR HR HU IE IL IS IT LI LT LU LV MT NL NO PL PT RO RS SE SI SK TR GB.

Scandinavia, for instance – the number of patents attributable to universities may be underestimated owing to the specific national rules on intellectual property attribution and patent ownership (the so-called "professor's privilege" in Sweden, Norway, Finland and some other countries[27] - see, among others: Hvide and Jones, 2018; Caviggioli et al., 2023a; Czarniztki et al., 2015; Lissoni et al., 2009). For completeness, we retain university units in these countries in the database, leaving the decision to exclude them to the analysis stage.

Second, we restrict the patenting offices considered. The original set of applications sourced from PATSTAT included those to all offices (APPLN_AUTH: all), but in fact the vast majority of patent applications worldwide originate in the five key patenting offices: according to WIPO around 85% of all filings in 2022 were at the IP offices of China, the US, Japan, the Republic of Korea and the EPO.[28] We thus analyse HEIs' knowledge creation according to patent applications filed in the IP5 jurisdictions: European Patent Office (EPO), Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO), China National Intellectual Property Administration (CNIPA) and United States Patent and Trademark Office (USPTO).[29] For purposes of comparison, however, we also provide a parallel series of indicators based exclusively on applications to EPO and USPTO. The indicators of patent quality, based on citation records from the OECD STI Micro-data Lab, are obtainable only for those two patent offices (see Section 2.4 for details).

The next choice was between all patent applications and patents granted only. Roughly half the university patents applied for in our dataset are eventually granted. This is in line with Davies et al., (2020), who also find that half of applications are successful; in Google Patents, the rate is 40%[30]. The patent granting process is long (on average, over five years for final approval), so retaining only patents granted would create a truncation problem at the end of the sample period. We thus use both types of data. That is, every indicator in the KC-HEI database is gauged both for all patent applications (e.g. the number of patent applications to IP5 in year *t* originating from a university *i* and attributable to CPC class *j*: $PA\_5\_fa_{ijt}$) and for granted patents only (e.g. the number of IP5 patents granted to an HEI in year *t* and CPC class *j*: $PA\_5\_fa\_g_{ijt}$). See Appendix A for the description of all the variables in KC-HEI.

As to the sample period, in line with Davies et al. (2023), who use the same edition of PATSTAT, we consider patent applications filed since 1980. Like all patent databases, PATSTAT

[27] "Professor's privilege" refers to the situation when university researchers enjoy full rights to their innovation (including patent ownership). Caviggioli et al. (2023a: 222) analyse patent filings between 1992 and 2014, providing a short summary of when and where the privilege was in force: "Sweden (in force throughout); Norway (until 2003), Germany (until 2001), Austria (until 2002), Finland (until 2007), Denmark (until 1999), Italy (from 2002)". They found underestimates owing to this phenomenon only for Sweden, Finland, and Norway. Hvide and Jones (2018) document a significant drop in both entrepreneurship and patenting rates by university researchers after the end of the "professor's privilege" in Norway. Similar results were found by Czarniztki et al. (2015) for Germany and Lissoni et al. (2009) for Denmark.

[28] Source: WIPO Statistics Database, March 2024 , https://www.wipo.int/en/ipfactsandfigures/patents [assessed on March 22, 2024].

[29] In PATSTAT TLS201_APPLN we choose the following entries in a*ppln_auth* variable: "EP", "US", "CN", "JP", "KR".

[30] 1790-2.24. https://patents.google.com/coverage Date of access: March 15, 2024.

suffers from end-of-period truncation bias, which depends on the considerable time that elapses between filing and the patent grant. We observe a significant decline in patent applications from 2020 onwards (Appendix C, Figure AppC_1), so we retain only the records from 1980 to 2019 in our dataset. Further, we are cautious in analysing the series based on patents granted and in using forward patent citations as indicators of patent quality or the "value" of the innovations patented (Trajtenberg, 1990; Verhoeven et al., 2016; Bloom and Van Reenen, 2002; Hall *et al.*, 2005; Gambardella *et al.*, 2008; Belenzon. 2012). As much as 5 or even 10 years after publication is needed for a patent to have a fair chance of being cited, so there is a significant citation lag (Gay et al., 2005; Marco, 2007; Munari and Oriani, 2011).[31]

All these restrictions leave a sample of 107,501 university patent applications to the IP5 since 1980, which can be used to derive the KC-HEI table.

2.3 The identification of HEIs in the patent data: the PATSTAT- ETER crosswalk

The biggest challenge for this procedure is identifying the patents of individual universities ensuring proper passage from raw patent-level records to university-level proxies of knowledge creation (patent counts). In line with Caviggioli et al. (2023b), we classify an application as a "university patent" if there is at least one HEI among the assignees. We take the data on patents of applicants, not inventors, because a university, as an institution, may appear among the patent contributors only as an applicant (while only a natural person can be an inventor). In allocating patents to university units, the benchmark analysis relies on fractional apportionment based on applicant shares within the patent application.

To identify university patents, we first use the information on the sector of activity provided by PATSTAT (variable *psn_sector* in TLS206_PERSON).[32] Specifically, we select such records in PATSTAT, for which *psn_sector* includes: UNIVERSITY. i.e. "UNIVERSITY". "GOV NON-PROFIT UNIVERSITY". "UNIVERSITY HOSPITAL". "COMPANY UNIVERSITY". The share of university patents in total PCT applications has risen slowly but steadily - from nearly zero in 1980 to almost 4.5% in 2000 and 6% in 2011 (Source: WIPO, 2011: 148).

However, one is left with a set of university patent applications identified by application number and the id's of assignees that are specific to the PATSTAT database (variable *psn_id*), making it impossible to analyse university-level determinants of knowledge creation, which would require merger

[31] For comparison, basic research takes an average almost 4 years to be cited in patent applications. Source: ERC Report: https://sciencebusiness.net/news-byte/basic-research-takes-average-37-years-to-be-cited-patent-applications [date of access: March 19, 2024).

[32] This variable is created using the additional file called PATSTAT Enhancements, constructed by ECOOM (Expertise Centre for Research and Development Monitoring), which provides sector allocation of applicants in PATSTAT. Researchers can apply for access to PATSTAT Enhancements data here: https://www.ecoom.be/en/data-collections/patstat-enhancements.

with external datasets (e.g. ETER) that use different identifiers (codes). This is a complex issue. A similar challenge is merging patent applicants in PATSTAT with firms from the Amadeus and Orbis databases (Bremer, 2023; Pompei & Venturini, 2022; Andrews at al., 2014). Some degree of harmonisation is provided by the OECD HAN Database, available also in PATSTAT (as *han_id* and *han_name* variables), but this solution is far from perfect (Caviggioli et al., 2023a, 2023b)[33].

We solve this problem by creating a crosswalk (available for research purposes in the replication package[34]) for correspondence between universities' names (and IDs) in PATSTAT (variable *psn_id*) and in ETER (variable *eter_id*). That is, we have created a tool that can make an immediate merge between PATSTAT data and all the other datasets using *eter_id* as university identifier (see Section 2.4).

The literature comprises only a few micro-level studies that use ETER data and patent statistics jointly. Many studies use the micro data at the level of HEIs, but actually perform the analysis at the regional level, because assigning university units to geographical regions is much easier than matching the names. For instance, Belvončíková (2021) identifies academic patents in OECD REGPAT by searching for the word "university" (in several linguistic mutations) in the assignees' names. This work combines the information from REGPAT and ETER but does not go beyond the regional level of analysis. Caviggioli et al. (2023a, 2023b) utilise PATSTAT and ETER, but they too remain at the regional level. Their sample is composed of the largest European universities, which run most EU-funded research projects. The authors underscore the problem of name harmonisation in PATSTAT (not entirely solved by the HAN harmonisation) and other issues, such as academic patents managed by technology transfer offices (TTOs) and ad-hoc companies related to the universities. The problem of name harmonisation is not specific to Europe - for instance, Squicciarini et al. (2012) employ both algorithm and manual procedures to deal with HEI name variations in the US data. One way to avoid a matching problem between patent data and university data is to collect the data from the various sources manually (as in, among others, Duarte et al., 2020; Acosta et al., 2012; Yamaguchi et al., 2019) or to carry out an own survey (as in, among others, Andersen and Rossi, 2011).However, these approaches usually result in a sample restricted to a single country or to a particular group of universities, selected *a priori*.

[33] These variables are sourced from OECD HAN Database (included in the OECD Patent datasets raw data, STI Micro-data Lab) and they provide some algorithm-based cleaning/harmonisation of assignees' names (a similar procedure is in place for *psn_id*). Still, one may find the same assignee under a few different *han_id*s (see also Graf and Menter, 2022). A similar problem occurs with the *psn_id*. Additionally, a *han_harmonized* variable is available in PATSTAT, which indicates whether the *han_name* could be matched with Orbis. However, it does not provide any ID code/number allowing for an easy match, so the matching still must be done using similarities in assignee's name, with the multiple problems of such a procedure (see Andrews at al., 2014; Pompei and Venturini, 2022).

[34] File *crosswalk_PSNid_to_ETERid.csv* . See the replication package.

Choosing among the many methods available (see Bremer, 2023), we elected to match the simplified names[35] of universities from ETER with the names of patent assignees in PATSTAT by means of string similarity metrics (the Levenshtein distance), followed by complex manual checks. A similar if smaller-scale approach is taken by Caviggioli et al. (2023b) and Squicciarini et al. (2012). The procedure is described in detail in Appendix B. The key issues addressed included: the presence of many "messy" names containing errors and redundant information, such as addresses and faculty names; university mergers or name changes, and the presence of patenting units (such as TTOs) that could be identified as HEI affiliates only by manual web search.

The identification of universities by an assignee code provided in PATSTAT (*psn_id*) is far from perfect, because a large number of different *psn_id*s may in fact refer to the same HEI. The crosswalk between PATSTAT and ETER that we constructed harmonises such multiple *psn_id*s with single *eter_id*s. This was a major problem - 80 or even more different *psn_id* codes could actually refer to one and the same university (this was the case, for instance, of the Federal Institute of Technology in Zurich and Trinity College Dublin). There were 182 HEIs with 10 or more different *psn_id*s. On average, 6 different *psn_id*s actually correspond to one *eter_id* (i.e. the same university).

Our success rate in matching PATSTAT with ETER is high: out of 7150 entities (*psn_ids*) attributed to the university sector in PATSTAT we identify, combine and match 6239 (i.e. 87%) with available *eter_id* identifiers. The crosswalk gives a unique correspondence between *psn_id* and *eter_id* for 1068 universities. Of these, we were able to match 866 European patenting HEIs (i.e. HEIs for which PATSTAT reports at least one patent application to IP5 in 1980-2019) with university records present in ETER, computing a microlevel dataset that can be used to detect the determinants of knowledge input provided by the university sector in Europe (Section 3).

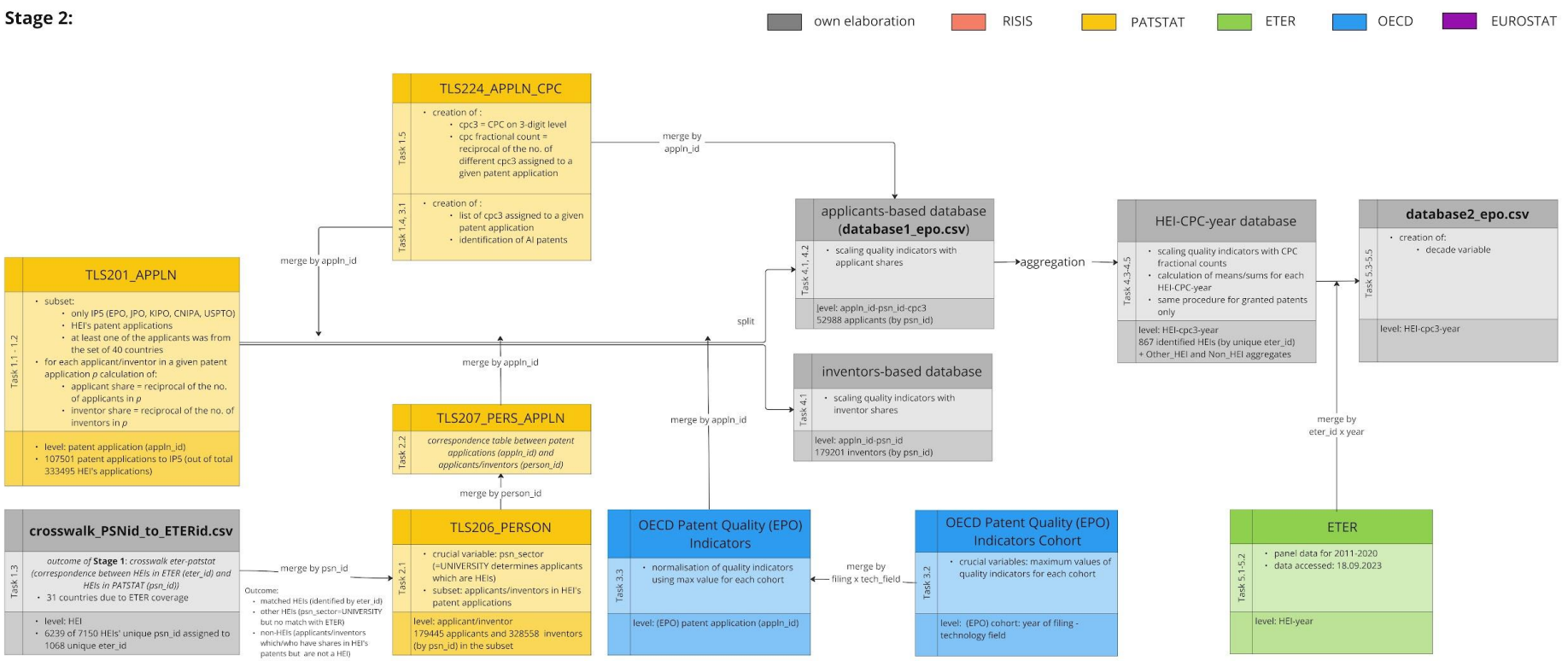
2.4 Database creation: intermediate steps

2.4.1 Derivation of KC-HEI (1980-2019)

The construction of KC-HEI is complex; the final database is derived from several intermediate ones that differ in level of aggregation. We begin with patent-level records, aggregate them to the level of universities and then to regions or countries. Figure 2 shows the main intermediate steps leading to the key KC-HEI (*database3*), as well as accompanying datasets (*database4* and *database5*), which we describe below.

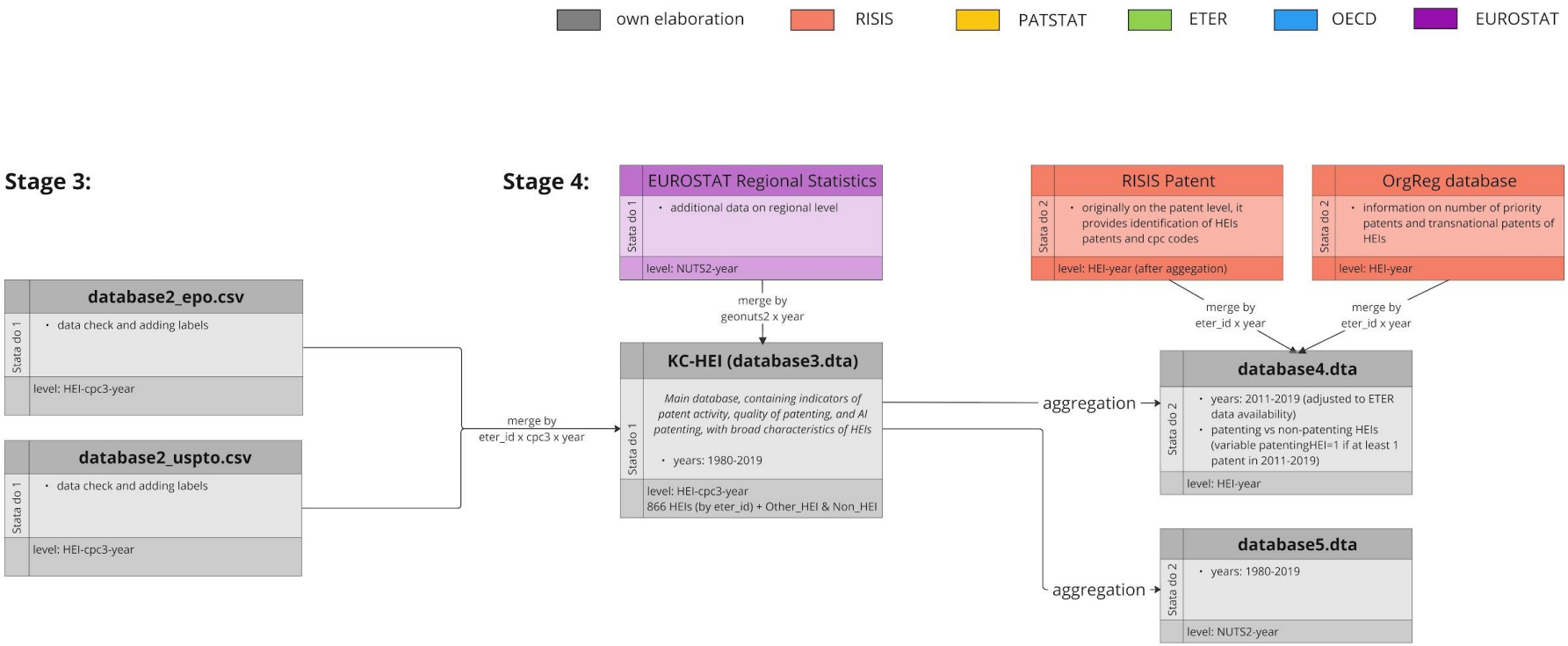
[35] Name simplification included the removal of language-specific characters, etc. (details in Appendix B).

Figure 2. The creation of the KC-HEI database and accompanying datasets - intermediate steps



Note: separately we use OECD Patent Quality Indicators and OECD Patent Quality Indicators Cohort for USTPO patents to generate **database1_uspto.csv** and **database2_uspto.csv** in a similar way.

Figure 2 (cont'd) The creation of KC-HEI database and accompanying datasets - intermediate steps



Source: own elaboration

The database designated *database1* consists of patent-level data, matching records on university patent applications from PATSTAT Global (Autumn 2022 edition) with a series of indicators capturing the technological and economic value of EPO and USPTO patents from the OECD Patent Quality Indicators Database (version: August 2023). We actually have two parallel versions of *database 1* depending on whether EPO or USPTO was used for its creation: both *database1_uspto.csv and database1_epo.csv* contain a total of 107,263 patent applications by 866 entities classified as universities.

In both versions of *database1* every observation is identified by the *appln_id-psn_id-cpc3* triad (see Section 2.2 for details). We allocate a portion of a patent application to HEI *i* using fractional apportionment by applicant shares: $\frac{1}{m}$ where *m* is the total number of applicants in a given patent application. Similarly, we allocate a portion of the overall patent that is "attributable" to the specific patented technology field (CPC 3 digit code): : $\frac{n_j}{\sum_k n_k}$, where n_j is the number of the most highly disaggregated CPC codes in the 3-digit CPC *j*. Therefore, for a given patent application *p* where HEI *i* is one of *m* applicants, the patenting activity of *i* in a CPC *j* equals the share $s_{ij}^p = \frac{1}{m}\frac{n_j}{\sum_k n_k}$. The fractional count procedure will better reflect the contribution of each HEI and avoid multiple counts of the same patent (Davies et al., 2023). However, for purposes of comparison we also keep simple patent counts in the dataset, i.e. the number of all patents in which a university was one of the applicants, regardless of the number of the latter (see below).

These two versions serve to derive two versions of *database2* (*database2_epo.csv* and *database2_uspto.csv*), in which the data is aggregated at the level of university units (i.e. the patent dimension is dropped and each row is identified by an HEI *i* -CPC3 *j* - year *t* triad). Finally, the two *database2s csv* are merged to derive *database3.csv,* our core university-level KC-HEI table, in which each record has a unique HEI identifier (*eter_id*) and name, CPC3 technological class, and year. KC-HEI contains several alternative indicators of universities' patenting activity, enriching the information in alternative datasets (OrgReg or RISIS Patents). A full list of the variables is given in Table 1A in Appendix A.

For instance, for a given HEI *i* we obtain the sum of fractional apportionments (shares) that it has in IP5 patent applications *p* attributable to a CPC technological class *j* as $\sum_p s_{ij}^p$. KC-HEI contains both fractional and simple counts (e.g. the number of all patent applications by technological field and year in which a given university was among the applicants, independent of the applicant share). Similarly, KC-HEI provides parallel series of indicators computed both with all patent applications and with only successful ones: for instance, the number of patent

applications to IP5 by HEI *i* in technological class *j* in year *t*, as well as the number of patents granted by *ijt*.

Further, for EPO and USPTO patents we also derive a set of indicators of patent quality (by fractional apportionment) based on the information on patent citations, claims, breakthroughness, and composite patent quality indices (Squicciarini et al., 2013; see Table A1 in Appendix A). Patent citations have long been used as an indicator of patent quality (at least since Trajtenberg, 1990); there is also a correlation between citation rates and the market value of a patent (Bloom and Van Reenen. 2002; Hall *et al.*, 2005; Gambardella *et al.*, 2008; Belenzon. 2012). In line with Squicciarini et al. (2013), we normalise quality indicators at individual patent level according to the corresponding maximum value of the patents in the same patent cohort (i.e. the combination of patent filing year and technology field). Then, the numbers are aggregated at the university level, either as a sum over all patents *p* of HEI *i*: $\sum_p \frac{QI}{maxQI} s_{ij}^p$ or as a mean: $\frac{1}{p}\sum_p \frac{QI}{maxQI} s_{ij}^p$, where: *QI* is a given quality indicator. and *maxQI* is the maximum value for this indicator for the cohort to which patent *p* belongs[36]. For each HEI we also provide the weighted number of so-called breakthrough patents (i.e. patents that are among the top 1% in citations): $\sum_p QI\ s_{ij}^p$ where: *QI* is a breakthrough indicator.

Additionally, in keeping with the development of research on the most recent wave of digital progress driven by AI technologies (including the literature on AI patents - among many, see Fujii and Managi, 2018; Giczy et al., 2022; Igna and Venturini, 2023; Parteka and Kordalska, 2023; Venturini, 2022; Balland and Boschma, 2021; Benassi et al., 2022; Czarnitzki et al., 2023; Yang, 2022), we identify AI patents in PATSTAT through AI cpc codes (WIPO) and construct indicators of AI knowledge creation by European universities (i.e. the number of AI university patents).

Detailed descriptions of all the variables in KC-HEI are in Appendix A - Table 1A. Table 1 reports the correlation coefficients between a selection of alternative university-level indicators present in KC-HEI: patent counts (based on all applications to IP5, EPO, USPTO; and patents granted), AI patents, selected quality-related measures (the number of breakthrough patents,

[36] The exceptions to these formulas include the following cases: not all the breakthroughness indicators are normalised (i.e. a patent can be classified as corresponding to a breakthrough innovation or not, 0-1 variable); the number of claims over the number of backward patent citations (*claims_bd*) is normalised using the maximum number of claims (*claims*), following Squicciarini et al. (2013). We use a sum for: number of backward patent citations (*bwd_cits* in Squicciarini et al., 2013), number of NPL citations (*npl_cits*), number of claims (*claims*), number of claims over number of backward patent citations (*claims_bwd*), all versions of variables regarding number of patent citations up to 5 years (f*wd_cits5, fwd.cits5_xy)* or 7 years (*fwd_cits7, fwd_cits7_*xy) after publication, and all versions of the breakthroughness indicator (*breakthrough, breakthrough_xy, breakthrough_x, breakthrough_y*). We use a mean for all the following indices: *generality, originality, radicalness, renewal*, and the composite quality indices (*quality_index_4, quality_index_6*).

forward citations, composite patent quality indices). It is evident that while the series of patent counts based on all applications and on patents granted only are highly correlated (0.94 in the case of IP5 patents), quality indicators offer a completely different type of information. For instance, the correlation between the number of patent applications filed by universities and the number of EPO breakthrough patents is just 0.27, while patent counts and composite indicators of universities' patent quality are practically uncorrelated. It implies that the choice of a particular indicator is likely to be critical in determining the conclusions drawn.

2.4.2 Accompanying datasets on patenting and non-patenting universities (2011-2019)

The data from KC-HEI (derived from PATSTAT), ETER and RISIS Patents is used to construct an additional dataset (*database 4*) on patenting and non-patenting HEIs, 2011-2019 (Sample 2). The shorter time period compared with *database 3* (KC-HEI) depends on the coverage of the ETER data, namely just 2011-2019. The division of HEIs into the patent-active/patent-inactive groups will be useful in assessing the determinants of activity/inactivity in knowledge creation, thanks to the presence of the control group of non-patenting universities (see Section 4).

We construct *database4* assuming that PATSTAT includes a universe of patenting entities. The most comprehensive set of patenting HEIs identifiable by a unique university code (*eter_id*) is then composed of universities present either in our KC-HEI table or in RISIS Patent/OrgReg (including 643 in both). After merging these two sources, we have 982 unique patenting HEIs (785 in KC-HEI plus 197 in RISIS Patent or OrgReg but not in KC-HEI). We then build a parallel sample of non-patenting HEIs that are identified as university units present in ETER but not in RISIS Patent/OrgReg or KC-HEI. In the end, this produces an additional set of microlevel data (*database4*) accompanying KC-HEI, consisting of 3054 universities (953 patenting and 2101 non-patenting) in 31 countries observed in 2011-2019; 1773 of the 3054 HEIs (58%) belong to the STEM group. Sample 2 consists solely of patenting HEIs for which we have PATSTAT-based indicators: 2886 HEIs, of which 785 patenting and 2101 non-patenting (2011-2019).

Finally, aggregating KC-HEI at the regional level, we construct *database5* with observations identified by *nuts2-year* pairs and providing regional-level indicators of university patenting activity in Europe (the leading regions in university patenting; see Section 3.4) merged with regional statistics at NUTS-2 level.

Table 1. Pairwise correlations between alternative indicators of university patenting activity in KC-HEI (Sample 1: 866 HEIs in 31 European countries, 1980-2019)

	Patent applications IP5 (FA*)	Granted patents IP5 (FA)	Patent applications EPO (FA*)	Patent applications USPTO (FA*)	AI patents IP5	Forward citations EPO patents	Forward citations USPTO patents	Breakthrough EPO patents	Breakthrough USPTO patents	EPO patents Quality (4components)	USPTO patents Quality (4 components)
Patent applications IP5 (FA*)	1.00										
Granted patents IP5 (FA)	0.84	1.00									
Patent applications EPO (FA*)	0.92	0.79	1.00								
Patent applications USPTO (FA*)	0.89	0.73	0.66	1.00							
AI patents IP	0.16	0.10	0.12	0.18	1.00						
Forward citations EPO patents	0.40	0.41	0.43	0.26	0.02	1.00					
Forward citations USPTO patents	0.25	0.24	0.16	0.31	0.01	0.10	1.00				
Breakthrough EPO patents	0.07	0.09	0.07	0.04	0.00	0.35	0.03	1.00			
Breakthrough USPTO patents	0.03	0.02	0.00	0.06	0.00	0.03	0.82	0.01	1.00		
EPO patents Quality (4components)	0.31	0.42	0.37	0.17	0.02	0.32	0.06	0.05	0.00	1.00	
USPTO patents Quality (4 components)	0.30	0.34	0.18	0.39	0.05	0.12	0.24	0.02	0.06	0.21	1.00

Note: FA - fractional apportionment. IP5, EPO, USPTO stand for patent offices: Breakthrough patents identified as top 1% cited patents. Indicators of patent quality based on OECD Patent Quality database (Squicciarini et al., 2013). All indicators described in Table 1A in Appendix A.
Source: own calculations using KC-HEI (Parteka et al., 2024)

2.5 Sample composition

The final KC-HEI database contains records on 866 HEIs in 31 European countries (listed in Table 2)[37] for the period 1980-2019 (unbalanced panel). The overlap between KC-HEI and ETER can be observed for the years 2011-2019. HEIs from Sample 1 account for almost 78% of all IP5 patent applications by universities present in PATSTAT Global (Autumn 2022) in the sample period and the sample countries.

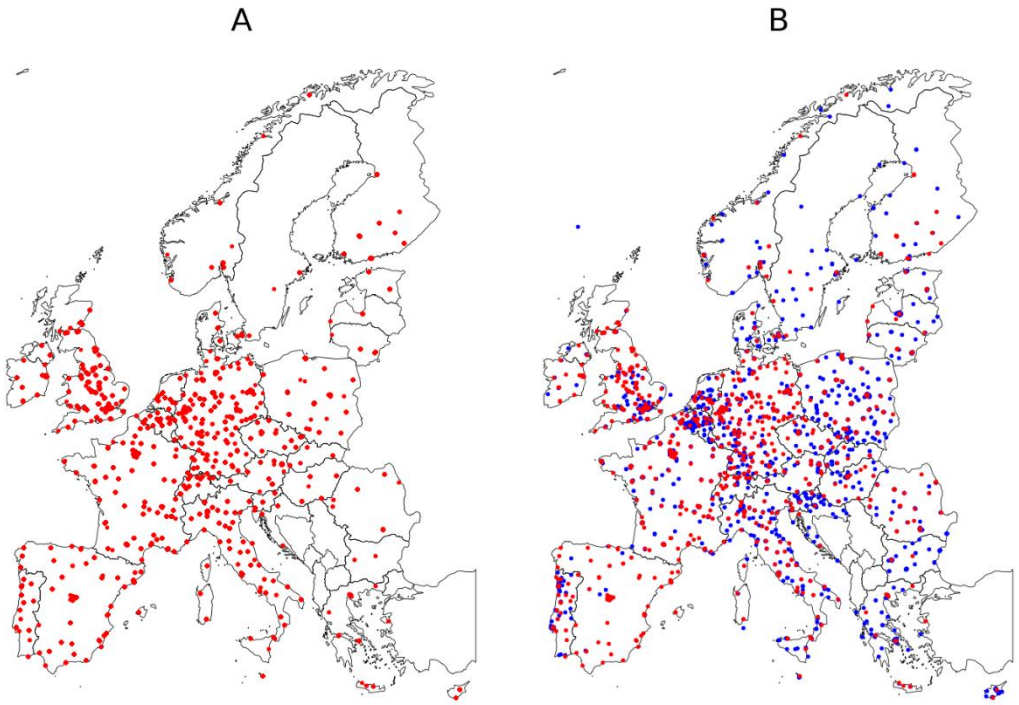


Figure 3. HEIs in the sample (red - patenting HEIs; blue - non patenting)

Note: Left map (Fig. 3A) - Sample 1: 866 patenting HEIs (1980-2019) in *database3* (KC-HEI), listed in file Sample1.xls in the replication package//Right map (Fig. 3B) - Sample 2: 785 patenting and 2101 non-patenting universities (2011-2019) in *database4* listed in file Sample2.xls in the replication package. Universities in small islands and overseas territories not shown.
Source: own elaboration using KC-HEI and accompanying datasets

We analyse universities throughout Europe (Figure 3); the number naturally varies from country to country. Table 1 reports the number of HEIs by country in the two sample periods, longer and shorter. In Sample 1 (only patenting HEIs, 1980-2019), most are located in Germany (167 universities, 19% of the entire sample), France (156, 18%) and the UK (117, 13%). In Sample 2 as well, which includes non-patenting institutions (a total of 2886 HEIs but a shorter time span: 2011-2019: Figure 3B), most are located in these same three countries. For the 31 countries in our sample, ETER contains information on 3070 HEIs (which can be taken as the

[37] AT, BE, BG, CH, CY CZ, DE, DK, EE, ES, FI, FR, GR, HR, HU, IE, IS, IT, LT, LU, LV, MT,NL, NO, PL, PT, RO, SE, SI, SK, UK.

most complete set of European HEIs on which institutional-level data can be gathered), so Sample 2 covers 94% of all European universities in the 31 countries.

Table 2. Sample composition - number (and share) of universities by country
(P- patenting* universities. NP** - non patenting universities)

		Sample 1: (KC-HEI: *database3*) 1980-2019		Sample 2: (*database 4*) 2011-2019					
		n=866 P		n=2886 (P+NP)		n=785 P		n=2101 NP	
Country code	Country name	number	share in Sample 1 [%]	number	share in Sample 2 [%]	number	share in Sample 2 [%]	number	share in Sample 2 [%]
AUT	Austria	25	2.89	71	2.5	22	2.80	49	2.33
BEL	Belgium	22	2.54	159	5.5	16	2.04	143	6.81
BGR	Bulgaria	2	0.23	51	1.8	2	0.25	49	2.33
CHE	Switzerland	21	2.42	38	1.3	20	2.55	18	0.86
CYP	Cyprus	2	0.23	31	1.1	2	0.25	29	1.38
CZE	Czech Republic	18	2.08	77	2.7	18	2.29	59	2.81
DEU	Germany	167	19.28	400	13.9	149	18.98	251	11.95
DNK	Denmark	7	0.81	43	1.5	6	0.76	37	1.76
ESP	Spain	64	7.39	79	2.7	62	7.90	17	0.81
EST	Estonia	5	0.58	29	1.0	3	0.38	26	1.24
FIN	Finland	17	1.96	47	1.6	16	2.04	31	1.48
FRA	France	156	18.01	392	13.6	145	18.47	247	11.76
GBR	UK	117	13.51	257	8.9	103	13.12	154	7.33
GRC	Greece	13	1.50	57	2.0	13	1.66	44	2.09
HRV	Croatia	3	0.35	42	1.5	3	0.38	39	1.86
HUN	Hungary	9	1.04	49	1.7	6	0.76	43	2.05
IRL	Ireland	18	2.08	26	0.9	17	2.17	9	0.43
ISL	Iceland	2	0.23	7	0.2	1	0.13	6	0.29
ITA	Italy	63	7.27	217	7.5	61	7.77	156	7.43
LTU	Lithuania	7	0.81	41	1.4	6	0.76	35	1.67
LUX	Luxembourg	1	0.12	3	0.1	1	0.13	2	0.10
LVA	Latvia	5	0.58	48	1.7	4	0.51	44	2.09
MLT	Malta	1	0.12	4	0.1	1	0.13	3	0.14
NLD	Netherlands	13	1.50	61	2.1	12	1.53	49	2.33
NOR	Norway	13	1.50	54	1.9	11	1.40	43	2.05
POL	Poland	50	5.77	284	9.8	42	5.35	242	11.52
PRT	Portugal	19	2.19	114	4.0	19	2.42	95	4.52
ROU	Romania	10	1.15	86	3.0	10	1.27	76	3.62
SVK	Slovakia	6	0.69	31	1.1	6	0.76	25	1.19
SVN	Slovenia	3	0.35	51	1.8	3	0.38	48	2.28
SWE	Sweden	7	0.81	37	1.3	5	0.64	32	1.52
Total	31 countries	866	100	2886	100	785	100	2101	100

Note: countries listed in alphabetical order (according to the country code)
*P - all the university units that were among assignees in at least one patent application in PATSTAT Global (Autumn 2022 Edition) in 2011-2019; ** NP - universities present in ETER (The European Tertiary Education Register - Lepori et al., 2023) but having no patent application filed in any of the three databases (KC-HEI, OrgReg, RISIS Patent) in 2011-2019

Source: own calculations

2.5 Matching KC-HEI with external datasets

One useful feature of KC-HEI is the possibility of merging it with many other datasets, making it an invaluable asset for research on the role of universities in the system of knowledge creation and the factors that shape that role.

The most complicated issue is microlevel matching between HEI-level patent statistics from PATSTAT and the characteristics of assignees (here: universities) in other, external datasets. Without a proper crosswalk, matching must be accomplished via string similarity, based on the names of entities[38] and/or other characteristics (addresses, say) when available in both datasets. This problem is solved by the crosswalk between applicant id's in PATSTAT and university id's in ETER (see Section 2.3 and Appendix B), so KC-HEI is easily merged with ETER. Using the crosswalk we can also check the coverage of KC-HEI against outside datasets that provide institution-level data on patents of European universities, such as the RISIS Patent database (Laurens, 2022)[39] and OrgReg Register (Lepori, 2022)[40]. The correlation between the number of university patents in RISIS Patent and KC-HEI is fairly strong (0.63, and rising to 0.74 when measured year-by-year)[41]. Similarly, the correlation between KC-HEI and OrgReg patenting comes to 0.60 (it too rising to 0.74 year-by-year). This is not a bad result considering the notable methodological differences in the construction of KC-HEI and these alternative datasets.[42]

Matching KC-HEI with country-level statistics is immediate, thanks to country codes. Similarly, university-level records can be matched with any regional database, such as that of Eurostat. using NUTS-2 or NUTS-3 region codes. Precise longitude and latitude for the location

[38] In firm-level studies this approach was taken by Andrews et al. (2014), who used string similarity to match ORBIS and PATSTAT. The authors note several problems with this procedure (for instance, probabilistic matching may produce some false negative or false positive patent assignments). Alternatively, one can match ORBIS Intellectual Property (IP), which links patents with companies, with PATSTAT (Dugoua and Gerarden, 2023). Lotti and Marin (2013) on AIDA data (by Bureau van Dijk) proposed a cleaning routine and several similarity scores for matching with PATSTAT. In their methodological paper the authors provided a literature review on previous attempts to match different datasets by Bureau van Dijk (such as ORBIS, AMADEUS, FAME) with PATSTAT. Pompei and Venturini (2022) applied string matching to combine ORBIS and REGPAT data, struggling with the same problems as in ORBIS-PATSTAT matching.

[39] RISIS Patents, also based on PATSTAT. covers 982 HEIs in 34 countries that applied for priority or transnational patents in the period 2000-2020. In RISIS applicant sectors are individual, company, unknown, government, non-profit, university and hospital. Access to RISIS Patent data was granted under the project: "Patents, technology and HEIs" (PATHEI).

[40] OrgReg reports data on organisations involved in research and higher education: higher education institutions, public research organisations, research hospitals, public administration entities, and private non-profit organisations.

[41] To calculate these correlations we take the variables that are closest in definitions and refer to the number of patent applications by universities: PA_5 in KC-HEI, all_patents ALL_patents in RISIS Patent and *Numberofprioritypatentapplic* in OrgReg.

[42] The total number of patents by HEIs in RISIS Patent is based on priority patents derived from the EPO PATSTAT (i.e. ipr_type = PI in PATSTAT), analogously in the OrgReg database of all patent offices produced by the European patent office (EPO), while KC-HEI uses information on all patent applications filed in the "big five" patent offices (IP5).

of HEIs allows the creation of maps (see Figure 1, maps in Section 3.4) or spatial analysis of patenting networks.

3. Stylised facts on European patenting using KC-HEI

Here we present descriptive statistics, maps, etc. that offer insights on the role of universities in European patenting and some key features of the system (such as its extreme concentration). Here we cannot demonstrate the full potential of KC-HEI. For instance, most of the evidence in this section uses the indicators for patents filed with IP5, but KC-HEI also provides parallel indices restricted to EPO or USPTO patents. Similarly, we focus here on the number of applications and patents granted, but KC-HEI also covers patent quality indicators and AI patents, which we leave for further exploration.

3.1. Trends over time

The number of university patents has increased over the years (Figure 4) – in keeping with the general upward trend in patent applications as recorded in PATSTAT (see Figure 1 and KIO - Davies et al. 2023, Figure 3). Appendix C shows the graph for the entire period up to 2022. It is clear that the apparent plunge in patent applications in 2020-2022 is an artificial result of the end-of-sample truncation in PATSTAT, which is why we end our analysis at 2019 (a similar approach has been adopted for the KIO database: Davies et al., 2023). A drop in the number of patents granted after 2015 mirrors the lag between application and patent granting. Finally, the information on university patents prior to 1990 is quite limited.

There is no great difference between the series generated with and those without fractional apportionment of patents (simple patent counts) - Figure 4. What strikes the eye, however, is the perceptible difference between IP5 applications and patents granted. The latter number is some 50% lower than that the former in 1980-2019[43] - that is, roughly half of the university patent applications are successful.

[43] Precisely, the proportion of grants to applications in Sample 1 is 47% in the series with fractional apportionment (FA) and 50% without it (simple patent counts).

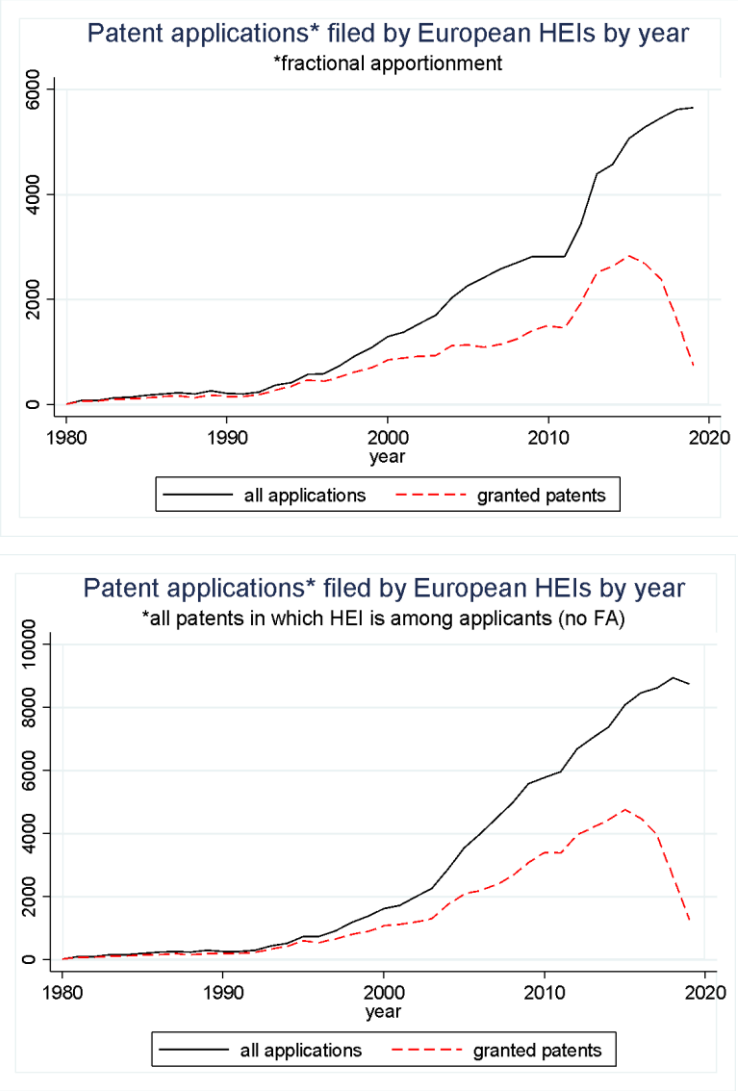


Figure 4. University patent applications filed by year (1980-2019)
Note: Sample of 866 patenting HEIs in 31 European countries (sample period only to 2019 to avoid end-of-sample truncation bias in PATSTAT). Patent applications filed with IP5.
Upper graph: patents allocated to HEIs using fractional apportionment (FA) by applicant share. Lower graph: no FA (patents allocated to HEIs independently of the applicant share)
Source: own calculations using KC-HEI database (Parteka et al., 2024)

3.2 Distribution of university patents by country

Using KC-HEI we find an extremely strong concentration of university patenting in Europe: just five countries account for over 70% of university patents. We arrive at this striking number by ranking the countries according to overall university patenting (Table 3). The unquestionable leaders are the UK, Germany, France, Switzerland, and Belgium – whose universities produced 72% of all university patent applications and 74% of patents granted to the 866 HEIs in our sample, 1980-2019. Universities in the leading country – the UK – account for a fourth of applications and patents (26% and 27%, respectively). The second-leading country (Germany) is responsible for 19% of applications and 18% of patents. Breakthrough inventions

(i.e. the top 1% patents by number of citations) come mainly from HEIs in Britain, Switzerland and Germany.

Of course, country size matters (some countries have many university units, others just a few - Table 2), so it is useful to recalculate to gauge number of patents per university (Table 4). In this case the small Benelux countries move up in the ranking.

3.2 Distribution of patents by university units

The concentration of university patenting can also be observed at the micro level. KC-HEI shows that a very few university units stand out as crucial nodes in the European innovation system (see Figure 1), while the bulk are totally absent from market-oriented innovation. The distribution of patenting activity across European universities, that is, is highly uneven: in KC-HEI Sample 1 of 866 patenting universities (thus excluding universities with no patents), the number of applications filed in 1980-2019 ranged from just 1 to over 2000 (summary statistics of key patent indicators by HEIs are reported in Table 2A in Appendix A). Figure 5 shows that 45% of the patenting universities have filed fewer than 10 applications to IP5 since 1980. At the other extreme, just 17 universities,[44] or 2% of the sample units, were among the assignees in more than 500 applications.

Sample 2, which contains data on both patenting and non-patenting universities (2011-2019), shows that a full 73% of European universities (2101 HEIs) were totally inactive (see Figure AppC_2 in Appendix C). In many countries the majority of HEIs never filed even a single patent application: Table 4 reports the percentage of patenting HEIs by country (Sample 2). For instance, in 2011-2019 half of Swiss universities filed at least one application, 40% of British, and a third of German universities, while in Bulgaria and Slovenia the figure was under 10%. These numbers should be treated with caution, however, because both the number of HEIs and their typology (technical universities tend to patent more) vary greatly between countries.

[44] University of Oxford, University College London, Federal Institute of Technology Lausanne, Federal Institute of Technology Zurich, Imperial College London, KU Leuven, University of Cambridge, Technical University of Denmark, Ghent University, University of Zurich, ULB, Karlsruhe Institute of Technology, Technical University of Munich, University of Manchester, University of Freiburg, Leiden University, Delft University of Technology.

Table 3. Ranking of countries in KC-HEI (1) by number of IP5 patent applications and patents. Top 5 countries in grey

Ranking by patent applications				Ranking by patents granted			
position	Country code	Country name	No. patent applications	position	Country code	Country name	No. patents
1	GBR	United Kingdom	14931.59	1	GBR	United Kingdom	7310.85
2	DEU	Germany	10495.73	2	DEU	Germany	4900.74
3	FRA	France	7564.01	3	FRA	France	4018.86
4	CHE	Switzerland	4384.33	4	CHE	Switzerland	1949.85
5	BEL	Belgium	3543.73	5	BEL	Belgium	1658.59
6	NLD	Netherlands	3063.62	6	NLD	Netherlands	1343.52
7	ITA	Italy	2416.07	7	ITA	Italy	1243.44
8	ESP	Spain	2245.52	8	ESP	Spain	986.70
9	DNK	Denmark	1678.51	9	IRL	Ireland	641.34
10	IRL	Ireland	1596.92	10	DNK	Denmark	637.33
11	AUT	Austria	1040.56	11	POL	Poland	469.65
12	POL	Poland	858.84	12	AUT	Austria	403.23
13	PRT	Portugal	525.06	13	CZE	Czech Republic	288.63
14	NOR	Norway	491.35	14	NOR	Norway	209.76
15	CZE	Czech Rep.	474.22	15	FIN	Finland	195.31
16	FIN	Finland	414.52	16	PRT	Portugal	191.00
17	HUN	Hungary	164.79	17	HUN	Hungary	92.31
18	LTU	Lithuania	108.06	18	SVN	Slovenia	55.88
19	SVN	Slovenia	102.46	19	LTU	Lithuania	52.56
20	EST	Estonia	92.23	20	EST	Estonia	45.59
21	LVA	Latvia	81.17	21	LVA	Latvia	36.33
22	SWE	Sweden	76.95	22	SWE	Sweden	23.34
23	LUX	Luxembourg	58.48	23	GRC	Greece	17.29
24	GRC	Greece	43.31	24	LUX	Luxembourg	15.77
25	ROU	Romania	22.85	25	HRV	Croatia	7.91
26	SVK	Slovakia	20.92	26	MLT	Malta	7.50
27	CYP	Cyprus	17.87	27	ROU	Romania	7.45
28	HRV	Croatia	14.66	28	CYP	Cyprus	7.37
29	MLT	Malta	12.75	29	ISL	Iceland	6.53
30	ISL	Iceland	10.37	30	SVK	Slovakia	6.33
31	BGR	Bulgaria	2.42	31	BGR	Bulgaria	0.75

Notes: Sample 1 - patenting HEIs in 31 countries, 1980-2019. Patents allocated to HEIs using fractional apportionment by applicant share.
Source: own calculations using KC-HEI database (Parteka et al., 2024)

Table 4. Ranking of countries in KC-HEI database (2) - patents per HEI

Top 5 countries in grey.

Ranking by number of patent applications per HEI				Ranking by patents granted per HEI			
position in the ranking	Country code	Country name	No. of patent applications per HEI	position in the ranking	Country code	Country name	No. of patents per HEI
1	DNK	Denmark	239.8	1	NLD	Netherlands	103.3
2	NLD	Netherlands	235.7	2	CHE	Switzerland	92.9
3	CHE	Switzerland	208.8	3	DNK	Denmark	91.0
4	BEL	Belgium	161.1	4	BEL	Belgium	75.4
5	GBR	United Kingdom	127.6	5	GBR	United Kingdom	62.5
6	IRL	Ireland	88.7	6	IRL	Ireland	35.6
7	DEU	Germany	62.8	7	DEU	Germany	29.3
8	LUX	Luxembourg	58.5	8	FRA	France	25.8
9	FRA	France	48.5	9	ITA	Italy	19.7
10	AUT	Austria	41.6	10	SVN	Slovenia	18.6
11	ITA	Italy	38.4	11	NOR	Norway	16.1
12	NOR	Norway	37.8	12	AUT	Austria	16.1
13	ESP	Spain	35.1	13	CZE	Czech Republic	16.0
14	SVN	Slovenia	34.2	14	LUX	Luxembourg	15.8
15	PRT	Portugal	27.6	15	ESP	Spain	15.4
16	CZE	Czech Republic	26.3	16	FIN	Finland	11.5
17	FIN	Finland	24.4	17	HUN	Hungary	10.3
18	EST	Estonia	18.4	18	PRT	Portugal	10.1
19	HUN	Hungary	18.3	19	POL	Poland	9.4
20	POL	Poland	17.2	20	EST	Estonia	9.1
21	LVA	Latvia	16.2	21	LTU	Lithuania	7.5
22	LTU	Lithuania	15.4	22	MLT	Malta	7.5
23	MLT	Malta	12.8	23	LVA	Latvia	7.3
24	SWE	Sweden	11.0	24	CYP	Cyprus	3.7
25	CYP	Cyprus	8.9	25	SWE	Sweden	3.3
26	ISL	Iceland	5.2	26	ISL	Iceland	3.3
27	HRV	Croatia	4.9	27	HRV	Croatia	2.6
28	SVK	Slovakia	3.5	28	GRC	Greece	1.3
29	GRC	Greece	3.3	29	SVK	Slovakia	1.1
30	ROU	Romania	2.3	30	ROU	Romania	0.7
31	BGR	Bulgaria	1.2	31	BGR	Bulgaria	0.4

Notes: Sample 1: patenting HEIs in 31 countries, 1980-2019. Patents allocated to HEIs using fractional apportionment by applicant share.

Source: own calculations using KC-HEI database (Parteka et al., 2024)

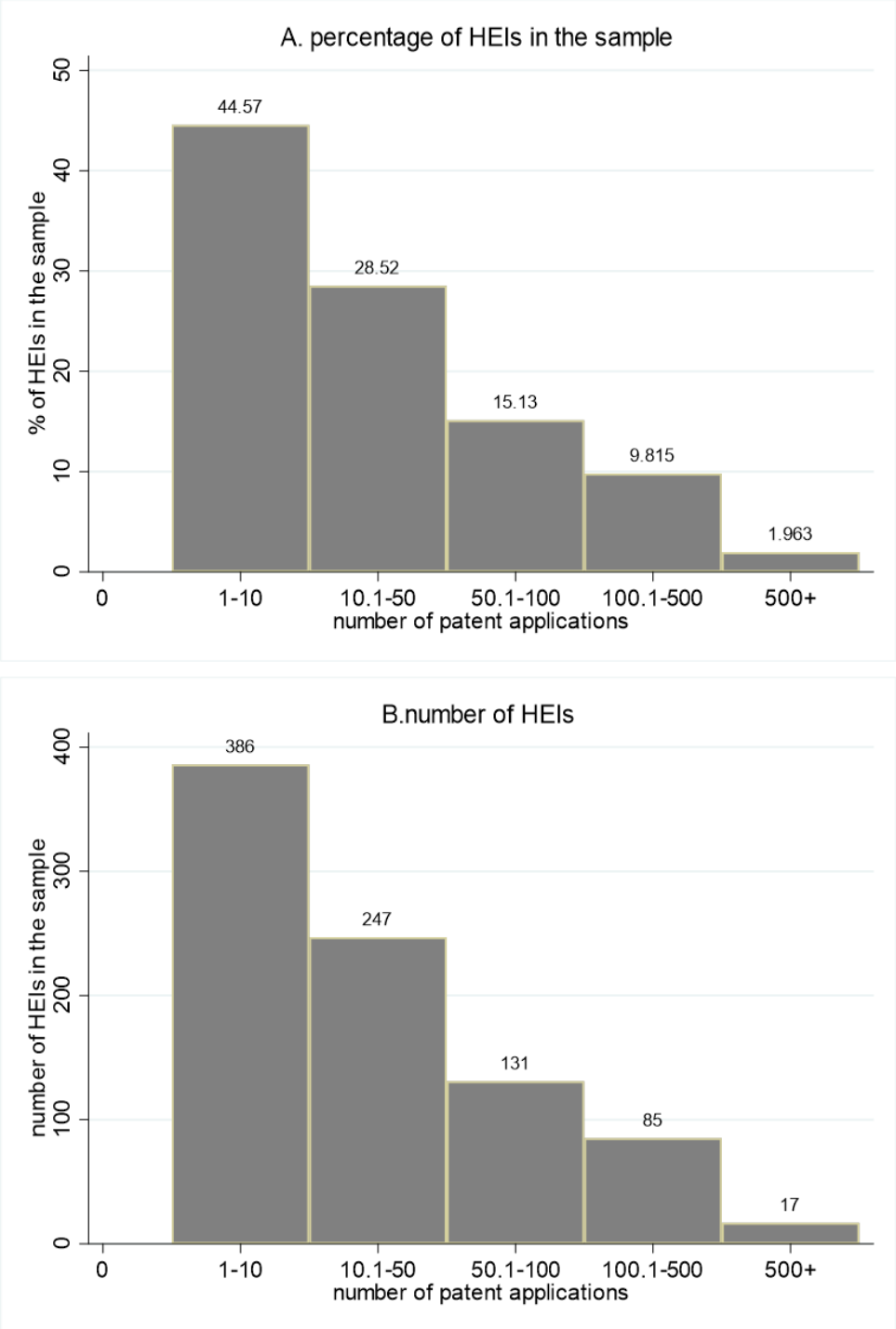


Figure 5. Distribution of patents among patenting HEIs (1980-2019)

Note: Sample of 866 patenting HEIs in 31 European countries (sample period only to 2019 to avoid end-of-sample truncation bias in PATSTAT). All patent applications filed with IP5. Patents allocated to HEIs using fractional apportionment (FA) by applicant share. Graph A – bins by percentage of HEIs in the sample. Graph B – bins by number of HEIs.

Source: own calculations using PATSTAT and KC-HEI database (Parteka et al., 2024)

Table 5. Percentage of patenting and non-patenting universities (all and STEM only), by country

		all HEIs		only STEM**	
Country code	Country name	number of universities in the sample	% of patenting* universities	number of STEM universities in the sample	% of patenting STEM universities
AUT	Austria	71	30.99	39	51.28
BEL	Belgium	159	10.06	47	21.28
BGR	Bulgaria	51	3.92	34	5.88
CHE	Switzerland	38	52.63	18	94.44
CYP	Cyprus	31	6.45	22	9.09
CZE	Czech Republic	77	23.38	33	51.52
DEU	Germany	400	37.25	237	62.03
DNK	Denmark	43	13.95	29	20.69
ESP	Spain	79	78.48	76	81.58
EST	Estonia	29	10.34	12	16.67
FIN	Finland	47	34.04	38	42.11
FRA	France	392	36.99	203	66.01
GBR	United Kingdom	257	40.08	158	62.66
GRC	Greece	57	22.81	46	28.26
HRV	Croatia	42	7.14	27	11.11
HUN	Hungary	49	12.24	28	21.43
IRL	Ireland	26	65.38	22	68.18
ISL	Iceland	7	14.29	6	16.67
ITA	Italy	217	28.11	80	73.75
LTU	Lithuania	41	14.63	28	21.43
LUX	Luxembourg	3	33.33	1	100.00
LVA	Latvia	48	8.33	25	16.00
MLT	Malta	4	25.00	2	50.00
NLD	Netherlands	61	19.67	38	31.58
NOR	Norway	54	20.37	36	30.56
POL	Poland	284	14.79	221	19.00
PRT	Portugal	114	16.67	64	29.69
ROU	Romania	86	11.63	9	33.33
SVK	Slovakia	31	19.35	18	33.33
SVN	Slovenia	51	5.88	12	25.00
SWE	Sweden	37	13.51	28	17.86
	TOTAL	2886	(mean) 27.209	1637	(mean) 45.45

Note: values based on Sample 2 (2011-2019)
*All the universities that were among assignees in at least one application in PATSTAT Global (Autumn 2022 edition) filed in 2011-2019. *Patenting HEI - at least one IP5 patent application in 1980-2019; **STEM institutions identified in line with ETER (Lepori, 2023: 119) as HEIs with a positive share of students and graduates in science and technology, i.e. in the fields 05, 06 and 07.
Source: own calculations

The dataset accompanying KC-HEI, *database4*, brings out the key differences between patenting and non-patenting HEIs in Europe (Table 6). Patenting universities are: older, much bigger, with lower student/teacher ratios, more active in terms of publishing, and wealthier, with higher core budgets[45] plus third-party funding.[46] The pattern among HEIs classified as STEM is similar.

Table 6. Characteristics of patenting and non-patenting universities - mean values

	All sample		STEM	
	patenting (n=785)	non-patenting (n=2101)	patenting (n=744)	non patenting (n=893)
Foundation year	1864.447	1971.27	1864.596	1977.959
Total number of academic staff (FTE)	1547.812	167.81	1567.232	237.374
Total number of students	19590.017	2940.495	19784.436	4772.95
Students per academic staff	15.037	21.115	14.973	25.349
Graduates per academic staff	3.692	5.263	3.675	6.321
Publication per academic staff	.281	.086	.278	.083
Total revenues (PPP)	2.554e+08	30488192	2.577e+08	43659772
Revenues per academic staff	174031.08	150916.16	172156.14	147452.02
Core revenues to total budget	.628	.587	.627	.614
Third party revenues to total budget	.182	.079	.181	.083

Note: values based on Sample 2 (2011-2019).
Source: own calculations using KC-HEI and accompanying datasets (Parteka et al., 2024).

3.4 Distribution of university patents and patent citations by region

University patenting in Europe displays a strong core-periphery pattern, matching regional/local economic divergences. The maps in Figure 6 show eight different categories of European NUTS2[47] regions - divided into groups according to university patenting percentiles (the top two classes, in red, correspond to the top 25% NUTS2 regions in number of patent applications - Figure 6A - left). The university innovation core is composed of university units from regions in the UK, southwestern Germany, Benelux, France, northern Italy and the regions where national capitals are located. The top 10 regions are: UKI3 (Inner London — West), FR10 (Ile-de-France), UKJ1(Berkshire, Buckinghamshire and Oxfordshire), CH04 (Zürich), CH01

[45] Core revenues are defined as funding available for the operations of the institution as a whole, not earmarked for specific activities, whose internal allocation can be decided fairly freely by the institution itself. In most institutions, the main component of the core budget is a government allocation (either national or regional). Source: Lepori (2023).

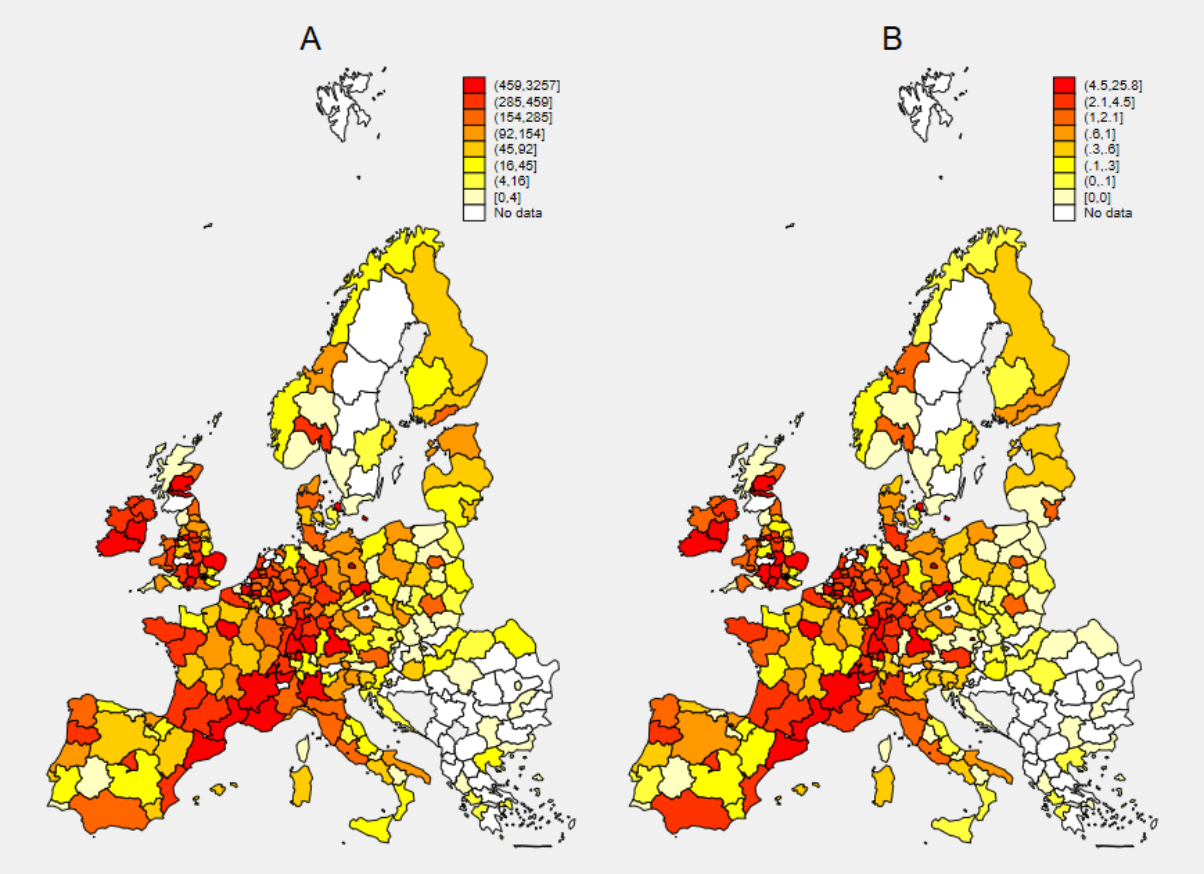
[46] Third-party funding is earmarked for specific activities and institutional units. It comprises: research grants from national and international funding agencies such as research councils (e.g. agencies like the Norwegian Research Council or the German DFG, European Union framework programmes, international programmes like Eureka or COST), funds from charities and non-profit organisations for specific research and educational purposes (like the Wellcome Trust of the Bill Gates foundation), contracts from public bodies, non-profit organisations and private companies for specific research and services and fees/payments from companies for educational services and research and service grants from companies. Source: Lepori (2023).

[47] KC-HEI allows for a similar analysis at the more detailed NUTS3 level. The concentration is very granular and often good performances by NUTS2 regions, in terms of university patenting, are actually driven by just a handful of HEIs (and thus more readily observable at NUTS 3 level). Results available upon request.

(Région lémanique), NL33 (Zuid-Holland), FRK2 (Rhône-Alpes), DK01 (Hovedstaden), DE12 (Karlsruhe), UKM7 (Eastern Scotland).

Figure 6. Regional distribution of university patents and citations, NUTS2 regions (1980-2019)

Figure 6A - patent applications; Figure 6B - patent citations [top regions in red]



Note: Sample of 866 patenting HEIs from 31 European countries. In Sweden, Norway and Finland the values are likely to be underestimated due to the "professor's privilege" - Caviggioli et al., 2023a, 2023b). Iceland, Cyprus and overseas regions not shown. Figure 6A: patent applications filed with IP5. Patents allocated to HEIs using fractional apportionment (FA) by applicant share. Figure 6B: patent citations measured by the number of forward citations of EPO patents in a 5-year period, normalised by the maximum number of citations in the corresponding technological field.

Legend: regions divided into 8 categories according to percentiles. Dark red colour indicates the top 12.5% regions in terms of university patenting/citations. Category 1 and 2 - top 25%; category 7 and 8 - bottom 25%. no data - regions without any HEIs or HEIs in the region have no patent activity.

Source: own calculations using KC-HEI database (Parteka et al., 2024)

Considering only patents granted, the picture does not change much[48]. For comparison, we show a similar map (Figure 6B - right) generated using patent citations (here: forward citations in a 5-year period, normalised by the maximum number of citations for each technological code and then summed at the level of HEIs and regions). This indicator can proxy for the real "quality" of patents (Squicciarini et al., 2013) and their effective impact in knowledge

[48] The additional maps/figures are available upon request.

flows (Davies et al., 2023). Mostly, cited patents come from HEIs in just a few top regions in Britain, Switzerland, France, Germany, Belgium and Denmark.

3.5 Trends by technology field. AI patents

KC-HEI allows analysis of university patents and their quality across 128 technological fields (listed in Table 3A in Appendix A). The top 10 fields in number of IP5 patent applications are listed in Table 7; Table 8 reports the top 10 cpc3 codes (more detailed). The leaders are "Human necessities" (among which "Medical or veterinary science; hygiene" is the top one) and "chemistry/metallurgy," which together account for more than half of all university patent applications.

Table 7. University patents by technological field

Ranking	Field code	Field name	univ. patent applications (IP5)		univ. granted patents (IP5)	
			total number	field share(%)	total number	field share (%)
1	A	Human necessities	15462.72	27.34	7033.80	26.21
2	C	Chemistry; metallurgy	14754.44	26.09	6493.90	24.20
3	G	Physics	11798.61	20.86	5601.12	20.88
4	H	Electricity	6243.07	11.04	3381.70	12.60
5	B	Performing operations; transporting	4721.15	8.35	2337.64	8.71
6	Y	Emerging Cross-Sectional Technologies	1668.03	2.95	948.48	3.53
7	F	Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	1185.32	2.10	636.10	2.37
8	E	Fixed constructions	401.90	0.71	228.39	0.85
9	D	Textiles; paper	318.60	0.56	170.58	0.64

Source: own calculation based on KC-HEI (Parteka et al., 2024)

Table 8. Top CPC3 technological codes in university patents (1980-2019)*

ranking	CPC3	Description of CPC3 code	Field code	Field description	CPC3 share (%)
1	A61	Medical or veterinary science; hygiene	A	Human necessities	24.90
2	G01	Measuring; testing	G	Physics	12.78
3	C12	Biochemistry; beer; spirits; wine; vinegar; microbiology; enzymology; mutation or genetic engineering	C	Chemistry; metallurgy	10.75
4	C07	Organic chemistry	C	Chemistry; metallurgy	8.23
5	H01	Electric elements	H	Electricity	6.03
6	G06	Computing; calculating or counting	G	Physics	3.54
7	B01	Physical or chemical processes or apparatus in general	B	Performing operations; transporting	3.44
8	H04	Electric communication technique	H	Electricity	2.66
9	C08	Organic macromolecular compounds; their preparation or chemical working-up; compositions based thereon	C	Chemistry; metallurgy	2.08
10	G02	Optics	G	Physics	1.96

Note: Top cpc3 - codes with the highest total number of patent applications by universities (fractional apportionment). *Sample 1: 866 HEIs, 1980-2019. Technological codes distributed within patent applications using code shares (one patent may be attributed to various cpc3 codes).
Source: own calculation based on KC-HEI (Parteka et al., 2024).

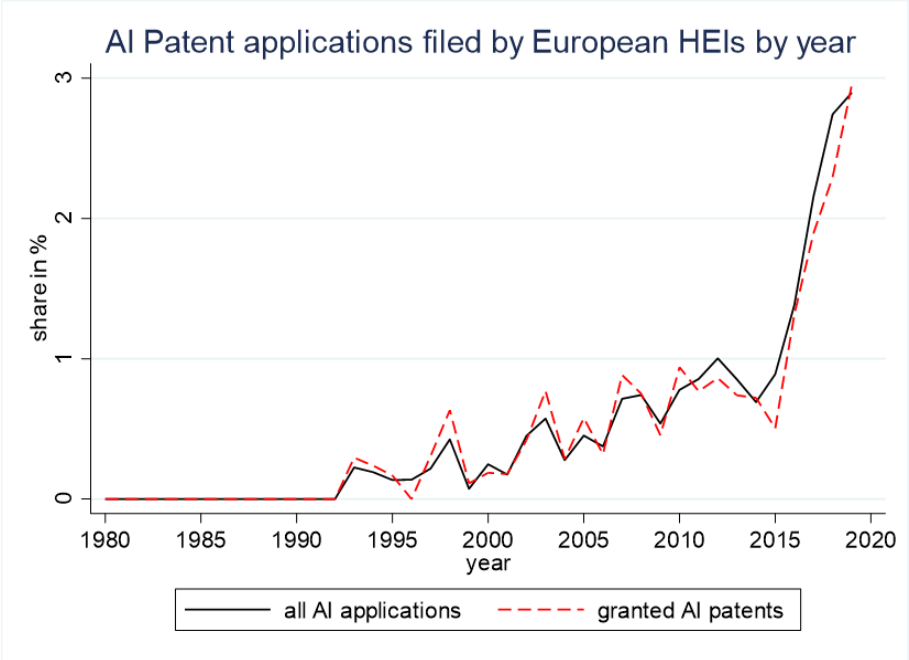


Figure 7. AI Patent applications filed by European HEIs (1980-2019)
Note: Sample of 866 patenting HEIs in 31 European countries. AI patents identified by AI CPC codes from WIPO (https://www.wipo.int/tech_trends/en/artificial_intelligence/patentscope.html)

KC-HEI can also document patterns of knowledge creation by universities in one particular field that has gained a great deal of attention in recent years: artificial intelligence. AI patenting by universities is limited, however. Only a very small proportion of university patents in 1980-2019 can be classified as AI-related (although this share did increase from zero in the 1980s to almost 3% of university patents in 2019). A full 75% of the HEIs in our sample are not at all active in AI patenting; only 25% (206 of 868 HEIs in Sample 1) applied for at least one AI-related patent in the period. Only 15% of the HEIs in the KC-HEI database were granted AI patents.

4. Conclusions

Do European universities play an important role in the European system of knowledge creation and market-oriented innovation? Which universities are the leaders in patenting, and how do they differ from those that do not patent their research? To help answer, we have constructed a new microlevel dataset (which we designate KC-HEI) covering almost 900 patenting universities in 31 European countries, plus an accompanying dataset on over 2100 non-patenting institutions of higher education. To construct KC-HEI we (i) scrutinised all the patent applications identifiable as university patents in PATSTAT (i.e. patent applications with at least one HEI among the applicants) since 1980 and (ii) created a unique crosswalk to match

university applicants in PATSTAT with other datasets containing university-level characteristics, such as ETER data.

Using KC-HEI, we document an extreme concentration of university patenting in Europe. Five countries account for over 70% of all patent applications by European universities. Of the universities that have filed at least one patent application since 1980, only 17 (2% of the units in our sample) were involved in more than 500 applications (gauged by fractional apportionment), while almost half had at most 10 IP5 patent applications in their records. Comparing patenting and non-patenting universities, an exercise that is possible only for a shorter time span (2011-2019), we find that three-fourths of European universities do not patent at all. Some have no need to patent, given their academic orientation to humanities and social sciences, but even among STEM institutions, which at least theoretically should be involved in applied innovation, about half have not made even a single IP5 patent application since 1980. Within countries, there are strong core-periphery patterns of university patenting at the NUTS 2 and NUTS 3 levels, tracking the regional distribution of per capita income in Europe.

The evidence produced here should be treated simply as one example of how the KC-HEI can be used; its actual potential in university-level studies on knowledge creation/patenting is much greater. KC-HEI contains an extensive series of patent-based variables that can be used to investigate such questions as: the role of HEIs in the European innovation system since 1980, their activity in various patent offices (IP5, EPO, USPTO), the quality of the patents generated by universities and their actual impact (patent citations, breakthrough innovations), differences across 128 technological fields or university patenting in the AI domain. Finally, the micro-level structure of the dataset allows detailed research on the determinants of innovation in higher education institutions, considering such characteristics as: HEI size, funding and employment structure, cooperation with firms and/or other universities, and location. Finally, using KC-HEI one can analyse the patenting performance of a given university vis-à-vis that typical of a broad European sample of university units and in response to changes in country-specific law on intellectual property or institution-specific reforms.

References

Acosta. M., Coronado. D., & Martínez. M. Á. (2012). Spatial differences in the quality of university patenting: Do regions matter?. *Research Policy*. 41(4). 692-703.

Agasisti, T., & Bertoletti, A. (2022). Higher education and economic growth: A longitudinal study of European regions 2000–2017. *Socio-Economic Planning Sciences*, *81*, 100940.

Aiello. F., Cardamone. P., Mannarino. L., & Pupo. V. (2022). Patents. family. and size: evidence from Italian manufacturing firms. *Economics of Innovation and New Technology*. https://doi.org/10.1080/10438599.2022.2134125

Andersen. B., & Rossi. F. (2011) UK universities look beyond the patent policy discourse in their intellectual property strategies. *Science and Public Policy* 38(4): 254–268. DOI: 10.3152/016502611X12849792159236.

Andrews D. Criscuolo C and Menon C (2014) Do Resources Flow to Patenting Firms?: Cross-Country Evidence from Firm Level Data. OECD Economics Department Working Papers 1127. Available at: http://dx.doi.org/10.1787/5jz2lpmk0gs6-en

Angori. G., Marzocchi. C., Ramaciotti. L., & Rizzo. U. (2023). A patent-based analysis of the evolution of basic. mission-oriented. and applied research in European universities. *The Journal of Technology Transfer*. https://doi.org/10.1007/s10961-023-10001-5

Ayerst. S., Ibrahim. F., MacKenzie. G., & Rachapalli. S. (2023). Trade and diffusion of embodied technology: an empirical analysis. *Journal of Monetary Economics*. *137*. 128-145.

Balland, P. A., & Boschma, R. (2021). Mapping the potentials of regions in Europe to contribute to new knowledge production in Industry 4.0 technologies. *Regional Studies*, *55*(10-11), 1652-1666.

Belenzon. S. (2012). Cumulative innovation and market value: Evidence from patent citations. *The Economic Journal. 122*(559). 265-285.

Behrens. V., & Trunschke. M. (2020). Industry 4.0 Related Innovation and Firm Growth. In ZEW Discussion Papers (No. 20–070; ZEW Discussion Papers). https://doi.org/10.2139/ssrn.3739871

Belvončíková. E. (2021) University owned patents in the EU28: EU15 versus EU13 spatial dimension. In Klímová. V., Žítek. V. (eds.) 24 th International Colloquium on Regional Sciences. Conference Proceedings. Brno: Masarykova univerzita. 2021. pp. 1–5. ISBN 978-80-210-9896-1.

Benassi, M., Grinza, E., Rentocchini, F., & Rondi, L. (2022). Patenting in 4IR technologies and firm performance. *Industrial and Corporate Change*, *31*(1), 112-136.

Bloom. N., & Van Reenen. J. (2002). Patents, real options and firm performance. *The Economic Journal. 112*(478). C97-C116.

Bilbao- Osorio. B., & Rodríguez- Pose. A. (2004). From R&D to innovation and economic growth in the EU. *Growth and Change*. 35(4). 434-455

Bremer. L. (2023). Fuzzy firm name matching: Merging Amadeus firm data to PATSTAT (No. 23-055/VIII). Tinbergen Institute. https://papers.tinbergen.nl/23055.pdf

Caviggioli. F., Colombelli. A., De Marco. A., Scellato. G., & Ughetto. E. (2023a). Co-evolution patterns of university patenting and technological specialization in European regions. *The Journal of Technology Transfe*r. 48(1). 216–239. https://doi.org/10.1007/s10961-021-09910-0

Caviggioli. F., Colombelli. A., De Marco. A., Scellato. G., & Ughetto. E. (2023b). The Impact of University Patenting on the Technological Specialization of European Regions: A Technology-Level Analysis. T*echnological Forecasting and Social Change*. 188. https://doi.org/10.2139/ssrn.4040647

Cesaroni. F., & Piccaluga. A. (2002). Patenting Activity of European Universities. Relevant? Growing? Useful?. In conference 'Rethinking Science Policy: Analytical Frameworks for Evidence-Based Policy (1-23).

Chalioti. E., Drivas. K., Kalyvitis. S., & Katsimi. M. (2020). Innovation. patents and trade: A firm-level analysis. *Canadian Journal of Economics*. 53(3). 949–981. https://doi.org/10.1111/caje.12451

Crespi. G., D'Este. P., Fontana. R., & Geuna. A. (2011). The impact of academic patenting on university research and its transfer. *Research Policy*. 40(1). 55-68.

Coupé. T. (2003). Science is golden: academic R&D and university patents. *The Journal of Technology Transfer*. 28(1). 31-46

Compagnucci. L., & Spigarelli. F. (2020). The Third Mission of the university: A systematic literature review on potentials and constraints. *Technological Forecasting and Social Change*. *161*. 120284

Conti. A., & Gaule. P. (2011). Is the US outperforming Europe in university technology licensing? A new perspective on the European Paradox. *Research Policy*. *40*(1). 123-135.

http://dx.doi.org/10.2139/ssrn.3633660

Czarnitzki. D., Doherr. T., Hussinger. K., Schliessler. P., & Toole. A. A. (2015). Individual versus institutional ownership of university-discovered inventions. Discussion Paper. Mannheim: ZEW.

Czarnitzki, D., Fernández, G. P., & Rammer, C. (2023). Artificial intelligence and firm-level productivity. *Journal of Economic Behavior & Organization*, *211*, 188-205.

Davies Ronald B., Dieter F. Kogler. and Ryan Hynes (2020). Patent Boxes and the Success Rate of Applications (2020). CESifo Working Paper No. 8375. http://dx.doi.org/10.2139/ssrn.3633660

Davies Ronald B., Dieter F. Kogler. Guohao Yang (2023). Construction of a Global Knowledge Input-Output Table. UCD Centre for Economic Research Working Paper Series WP23/22. https://www.ucd.ie/economics/t4media/WP23_22.pdf

Dosi, G., Llerena, P., & Labini, M. S. (2006). The relationships between science, technologies and their industrial exploitation: An illustration through the myths and realities of the so-called 'European Paradox'. *Research policy*, *35*(10), 1450-1464.

Duarte MGP. Gonçalves E. Chein F. et al. (2020) Drivers of scientific-technological production in brazilian higher education and research institutions. *Revista de Economia Contemporanea* 24(3): 1–41. DOI: 10.1590/198055272432.

Dugoua, E., & Gerarden, T. (2023). *Induced Innovation, Inventors, and the Energy Transition* (No. w31714). National Bureau of Economic Research.

Etzkowitz. H., & Leydesdorff. L. (1997). Introduction to special issue on science policy dimensions of the Triple Helix of university-industry-government relations. *Science and Public Policy*. *24*(1). 2-5.

EC (1995) Green Paper on Innovation. European Commission. https://op.europa.eu/en/publication-detail/-/publication/ad1d6f21-0b2e-423f-9301-c608035e906f

EC (2022a). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT. THE COUNCIL. THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on a European strategy for universities. European Commission. https://education.ec.europa.eu/sites/default/files/2022-01/communication-european-strategy-for-universities.pdf [date of access: March 13. 2024]

EC (2022b) The management and commercialisation of intellectual property in European universities. European Commision.

https://op.europa.eu/en/publication-detail/-/publication/cfa4b8b0-99d7-11ec-83e1-01aa75ed71a1/language-en

Fujii. H., & Managi. S. (2018). Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. *Economic Analysis and Policy*. *58*. 60-69.

Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. *European Management Review*, *5*(2), 69-84.

Gay. C., Le Bas. C., Patel. P., & Touach. K. (2005). The determinants of patent citations: An empirical analysis of French and British patents in the US. *Economics of Innovation and New Technology. 14*(5). 339–350.

Giczy. A. V., Pairolero. N. A., & Toole. A. A. (2022). Identifying artificial intelligence (AI) invention: A novel AI patent dataset. *The Journal of Technology Transfer*. *47*(2). 476-505.

Graf H. and Menter M. (2022) Public research and the quality of inventions: the role and impact of entrepreneurial universities and regional network embeddedness. *Small Business Economics* 58(2). Springer: 1187–1204. DOI: 10.1007/s11187-021-00465-w.

Gurmu. S., Black. G. C., & Stephan. P. E. (2010). The knowledge production function for university patenting. *Economic Inquiry.* 48(1). 192-213.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, *36*(1), 16-39.

Henderson R. Jaffe AB and Trajtenberg M (1998) Universities as a source of commercial technology: A detailed analysis of university Patenting. 1965-1988. *Review of Economics and Statistics* 80(1): 119–127. DOI: 10.1162/003465398557221.

Hvide. H. K., & Jones. B. F. (2018). University innovation and the professor's privilege. *American Economic Review*. 108(7). 1860-1898

Igna. I., & Venturini. F. (2023). The determinants of AI innovation across European firms. *Research Policy*. *52*(2). 104661.

Laurens P. (2022). DOCUMENTATION OF RISIS DATASETS: RISIS Patent Database. LISIS. Univ Gustave Eiffel. ESIEE Paris. CNRS. INRAE. hal-03881028

Lepori B. (2022) OrgReg Methodological manual. Zenodo. https://doi.org/10.5281/zenodo.6396703

Lepori B. (2023) European Tertiary Education Register (ETER) Handbook for Data Collection [ETER Handbook]. https://zenodo.org/records/10065818

Lepori. B., Lambrechts. A.A., Wagner-Schuster. D. *et al.* (2023). The European Tertiary Education Register. the reference dataset on European Higher Education Institutions. *Sci Data* 10. 438 https://doi.org/10.1038/s41597-023-02353-2

Lissoni F. Lotz P. Schovsbo J. et al. (2009) Academic patenting and the professor's privilege: Evidence on Denmark from the KEINS database. *Science and Public Policy* 36(8): 595–607. DOI: 10.3152/030234209X475443.

Lotti, F. and Marin, G. 2013. Matching of PATSTAT applications to AIDA firms - Discussion of the methodology and results. *Occasional Papers* No. 166, Bank of Italy.

Marco. A. C. (2007). The dynamics of patent citations. *Economics Letters. 94*(2). 290–296

Munari, F., & Oriani, R. (Eds.). (2011). *The economic valuation of patents: Methods and applications*. Cheltenham: Edward Elgar Publishing.

Nagaoka. S., Motohashi. K., & Goto. A. (2010). Patent statistics as an innovation indicator. In *Handbook of the Economics of Innovation* (Vol. 2. pp. 1083-1127). North-Holland.

OECD. REGPAT database. August 2022.

Parteka, A., & Kordalska, A. (2023). Artificial intelligence and productivity: global evidence from AI patent and bibliometric data. *Technovation*, *125*, 102764.

Pompei, F. and Venturini, F. (2022), "Firm level productivity and profitability effects of managerial and organisational capabilities and innovations", Untangled Research Papers No. 02/2022, available at: https://projectuntangled.eu/untangled-research-papers/

Rodríguez-Navarro. A., & Narin. F. (2018). European paradox or delusion—are European science and economy outdated?. *Science and Public Policy*. *45*(1). 14-23.

Squicciarini. M., Millot. V., & Dernis. H. (2012). Universities' trademark patterns and possible determinants. *Economics of Innovation and New Technology*. 21(5–6). 473–504. https://doi.org/10.1080/10438599.2012.656526

Squicciarini. M., H. Dernis and C. Criscuolo (2013). Measuring Patent Quality: Indicators of Technological and Economic Value. OECD Science. Technology and Industry Working Papers. 2013/03. OECD Publishing. http://dx.doi.org/10.1787/5k4522wkw1r8-en

Trajtenberg. M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand journal of economics*. 172-187.

Verhoeven. D., Bakker. J., & Veugelers. R. (2016). Measuring technological novelty with patent-based indicators. *Research policy*. *45*(3). 707-723.

Whalley A and Hicks J (2014) SPENDING WISELY? HOW RESOURCES AFFECT KNOWLEDGE PRODUCTION IN UNIVERSITIES. *Economic Inquiry* 52(1). John Wiley & Sons. Ltd: 35–55. DOI: 10.1111/ECIN.12011.

WIPO (2011). World Intellectual Property Report. The Changing Face of Innovation https://www.wipo.int/edocs/pubdocs/en/intproperty/944/wipo_pub_944_2011.pdf [date of access: April 15, 2024]

WIPO (2023). Patent Cooperation Treaty Yearly Review 2023 The International Patent System. https://www.wipo.int/edocs/pubdocs/en/wipo-pub-901-2023-en-patent-cooperation-treaty-yearly-review-2023.pdf [date of access: March 13, 2024]

Yamaguchi. Y., Fujimoto. J., Yamazaki. A., & Koshiyama. T. (2019). Analysis of the factors influencing patent creation and patent-based technology transfer in universities. In 2019 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-10). IEEE

Yang, C. H. (2022). How artificial intelligence technology affects productivity and employment: firm-level evidence from Taiwan. *Research Policy*, *51*(6), 104536.

Appendix A. Details on KC-HEI dataset

Table 1A. List of variables in KC-HEI and corresponding data sources

group	details	variable name	variable description	source
identification	**HEIs names & codes**	HEI_name	HEI name as in ETER. For Other_HEI: "all other patenting HEIs not identified as single units"; for Non_HEI : "all other patenting units not being HEIs"	ETER: BAS.INSTNAME.Text
		HEI_name_eng	HEI English name as in ETER. For Other_HEI: "all other patenting HEIs not identified as single units"; for Non_HEI: "all other patenting units not being HEIs"	ETER: BAS.INSTNAMEENGL.Text
		eter_id	HEI id as in ETER for matched units; else either "Other_HEI" for unmatched HEIs, or "Non_HEI" for all other patenting units not being HEIs	ETER Crosswalk file (matching psn_id in PATSTAT with eter_id): crosswalk_PSNid_to_ETERid.csv Details in Appendix B.
		HEI	binary variable =1 for HEI (identified by eter_id or marked as Other_HEI); =0 else (applicants marked as Non_HEI)	own calculations, based on eter_id
	geo/location	country	3 digit ISO country code where HEI is located [corresponding with cited_country / citing_country in KIO (Davies et al., 2023)]	own, in accordance with country_2g
		country_2g	2 digit ISO country code (except for United Kingdom here coded as UK) where HEI is located	ETER: two first letters of eter_id
		country_name	Country name where HEI is located	own, in accordance with country_2g
		geonuts2	NUTS 2 region of establishment	ETER: GEO.NUTS2
		geonuts3	NUTS 3 region of establishment	ETER: GEO.NUTS3
		geocity	Name of the city	ETER: GEO.CITY
		cpc3	3 digit CPC code [corresponding with cited_cpc3 / citing_cpc3 in KIO (Davies et al., 2023)]	extracted from PATSTAT: cpc_class_symbol in table TLS224_APPLN_CPC

Time dimension		year	year corresponding to patent application filing year	PATSTAT: appln_filing_year ETER: *BAS.REFYEAR.Year* OECD Patent Quality database*: filing*
		decade	=1980s for 1980-89; =1990s for 1990-99; =2000s for 2000-2009; =2010s for 2010-2019 [corresponding with cited_decade /citing_decade in KIO (Davies et al., 2023)]	own calculations, based on year
Indicators of patenting activity	**fractional apportionment (fa)**	PA_5_fa	the number of patent applications (by applicant) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; fractional apportionment (using applicant share) and CPC fractional count	Own calculations based on PATSTAT
		PA_EPO_fa	the number of patent applications (by applicant) to EPO; fractional apportionment (using applicant share) and CPC fractional count	Own calculations based on PATSTAT
		PA_USPTO_fa	the number of patent applications (by applicant) to USPTO; fractional apportionment (using applicant share) and CPC fractional count	Own calculations based on PATSTAT
	fractional apportionment (fa), granted patents only (g)	PA_5_fa_g	the number of patent applications (by applicant) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; fractional apportionment (using applicant share) and CPC fractional count, only granted patents	Own calculations based on PATSTAT
		PA_EPO_fa_g	the number of patent applications (by applicant) to EPO; fractional apportionment (using applicant share) and CPC fractional count, only granted patents	Own calculations based on PATSTAT
		PA_USPTO_fa_g	the number of patent applications (by applicant) to USPTO; fractional apportionment (using applicant share) and CPC fractional count, only granted patents	Own calculations based on PATSTAT
	simple counts - no fractional apportionment (all patent applications in	PA_5	the number of patent applications (by applicant) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; no fa, CPC fractional count	Own calculations based on PATSTAT

	which HEI is among applicants)	PA_5_AI	Identification of AI patents (non-zero value identifies AI patent in HEI-CPC-year database; in the HEI-year database this is the number of AI patent applications (by HEI) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; no fa)	Own calculations based on PATSTAT and AI identifiers from PATENTSCOPE Artificial Intelligence Index from WIPO
	simple counts - no fractional apportionment (all patent applications in which HEI is among applicants), granted patents only (g)	PA_5_g	the number of patent applications (by applicant) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; no fa, CPC fractional count, only granted patents	Own calculations based on PATSTAT
		PA_5_AI_g	Identification of AI patents (non-zero value identifies AI patent in HEI-CPC-year database; in the HEI-year database this is the number of AI patent applications (by HEI) to 5 patent offices: EPO, USPTO, CPO, JPO, KPO; no fa), only granted patents	Own calculations based on PATSTAT and AI identifiers from PATENTSCOPE Artificial Intelligence Index from WIPO
Normalised indicators of patenting activity (patent quality indicators)		BC_EPO_fa_ns	sum of Backward citations of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: bwd_cits) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		NPLC_EPO_fa_ns	sum of Citations to non-patent literature of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: npl_cits) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		CL_EPO_fa_ns	sum of Patent claims of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: claims) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		CLb_EPO_fa_ns	sum of Backwards patent claims of patent applications (by applicant) to EPO; normalised using max for cohort (same as for CL_EPO_fa_ns); multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: claims_bwd) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC5_EPO_fa_ns	sum of Forward citations in 5-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits5) and OECD Patent Quality (EPO)

				Indicators Cohort (for normalisation)
		fC5xy_EPO_fa_ns	sum of Forward citations XY in 5-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits5_xy) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC7_EPO_fa_ns	sum of Forward citations in 7-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits7) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC7xy_EPO_fa_ns	sum of Forward citations XY in 7-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits7_xy) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		BT_EPO_fa_s	sum of Breakthroughness of patent applications (by applicant) to EPO; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: breakthrough)
		BTxy_EPO_fa_s	sum of Breakthroughness XY of patent applications (by applicant) to EPO; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: breakthrough_xy)
		G_EPO_fa_nm	mean of Generality index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: generality) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		O_EPO_fa_nm	mean of Originality index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: originality) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		RD_EPO_fa_nm	mean of Radicalness index of patent applications (by applicant) to EPO; normalised using max for cohort;	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: radicalness) and OECD Patent Quality

			multiplied by PA_EPO_fa	(EPO) Indicators Cohort (for normalisation)
		RN_EPO_fa_nm	mean of Patent renewal of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: renewal) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		QI4_EPO_fa_nm	mean of Quality index (4 components) of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: quality_index_4) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		QI6_EPO_fa_nm	mean of Quality index (6 components) of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: quality_index_6) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		BC_USPTO_fa_ns	sum of Backward citations of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: bwd_cits) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		NPLC_USPTO _fa_ns	sum of Citations to non-patent literature of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: npl_cits) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		CL_USPTO _fa_ns	sum of Patent claims of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: claims) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		CLb_USPTO _fa_ns	sum of Backwards patent claims of patent applications (by applicant) to USPTO; normalised using max for cohort (same as for CL_USPTO_fa_ns); multiplied by	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: claims_bwd) and OECD Patent Quality (USPTO) Indicators Cohort (for

			PA_USPTO_fa	normalisation)
		fC5_USPTO _fa_ns	sum of Forward citations in 5-years period of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: fwd_cits5) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		fC7_USPTO _fa_ns	sum of Forward citations in 7-years period of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: fwd_cits7) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		BTx_USPTO _fa_s	sum of Breakthroughness X of patent applications (by applicant) to USPTO; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: breakthrough_x)
		BTy_USPTO _fa_s	sum of Breakthroughness Y of patent applications (by applicant) to USPTO; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: breakthrough_y)
		G_USPTO _fa_nm	mean of Generality index of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: generality) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		O_USPTO _fa_nm	mean of Originality index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: originality) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		RD_USPTO _fa_nm	mean of Radicalness index of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: radicalness) and OECD Patent Quality (USPTO) Indicators Cohort (for

				normalisation)
		RN_USPTO _fa_nm	mean of Patent renewal of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: renewal) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		QI4_USPTO _fa_nm	mean of Quality index (4 components) of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: quality_index_4) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		QI6_USPTO _fa_nm	mean of Quality index (6 components) of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: quality_index_6) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
	granted patents only (g)	BC_EPO_fa_ns_g	sum of Backward citations of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: bwd_cits) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		NPLC_EPO_fa_ns_g	sum of Citations to non-patent literature of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: npl_cits) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		CL_EPO_fa_ns_g	sum of Patent claims of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: claims) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		CLb_EPO_fa_ns_g	sum of Backwards patent claims of patent applications (by applicant) to EPO; normalised using max for cohort (same as for CL_EPO_fa_ns); multiplied by	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: claims_bwd) and OECD Patent Quality

			PA_EPO_fa; only granted patents	(EPO) Indicators Cohort (for normalisation)
		fC5_EPO_fa_ns_g	sum of Forward citations in 5-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits5) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC5xy_EPO_fa_ns_g	sum of Forward citations XY in 5-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits5_xy) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC7_EPO_fa_ns_g	sum of Forward citations in 7-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits7) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		fC7xy_EPO_fa_ns_g	sum of Forward citations XY in 7-years period of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: fwd_cits7_xy) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		BT_EPO_fa_s_g	sum of Breakthroughness of patent applications (by applicant) to EPO; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: breakthrough)
		BTxy_EPO_fa_s_g	sum of Breakthroughness XY of patent applications (by applicant) to EPO; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: breakthrough_xy)
		G_EPO_fa_nm_g	mean of Generality index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: generality) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		O_EPO_fa_nm_g	mean of Originality index of patent applications (by applicant) to EPO; normalised using max for cohort;	Own calculations based on OECD Patent Quality (EPO) Indicators (variable:

			multiplied by PA_EPO_fa; only granted patents	originality) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		RD_EPO_fa_nm_g	mean of Radicalness index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: radicalness) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		RN_EPO_fa_nm_g	mean of Patent renewal of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: renewal) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		QI4_EPO_fa_nm_g	mean of Quality index (4 components) of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: quality_index_4) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		QI6_EPO_fa_nm_g	mean of Quality index (6 components) of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_EPO_fa; only granted patents	Own calculations based on OECD Patent Quality (EPO) Indicators (variable: quality_index_6) and OECD Patent Quality (EPO) Indicators Cohort (for normalisation)
		BC_USPTO_fa_ns_g	sum of Backward citations of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: bwd_cits) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		NPLC_USPTO _fa_ns_g	sum of Citations to non-patent literature of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: npl_cits) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		CL_USPTO _fa_ns_g	sum of Patent claims of patent applications (by applicant) to USPTO; normalised using max for cohort;	Own calculations based on OECD Patent

			multiplied by PA_USPTO_fa; only granted patents	Quality (USPTO) Indicators (variable: claims) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		CLb_USPTO _fa_ns_g	sum of Backwards patent claims of patent applications (by applicant) to USPTO; normalised using max for cohort (same as for CL_USPTO_fa_ns); multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: claims_bwd) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		fC5_USPTO _fa_ns_g	sum of Forward citations in 5-years period of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: fwd_cits5) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		fC7_USPTO _fa_ns_g	sum of Forward citations in 7-years period of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: fwd_cits7) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		BTx_USPTO _fa_s_g	sum of Breakthroughness X of patent applications (by applicant) to USPTO; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: breakthrough_x)
		BTy_USPTO _fa_s_g	sum of Breakthroughness Y of patent applications (by applicant) to USPTO; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: breakthrough_y)
		G_USPTO _fa_nm_g	mean of Generality index of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: generality) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)

		O_USPTO _fa_nm_g	mean of Originality index of patent applications (by applicant) to EPO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: originality) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		RD_USPTO _fa_nm_g	mean of Radicalness index of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: radicalness) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		RN_USPTO _fa_nm_g	mean of Patent renewal of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: renewal) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		QI4_USPTO _fa_nm_g	mean of Quality index (4 components) of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: quality_index_4) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)
		QI6_USPTO _fa_nm_g	mean of Quality index (6 components) of patent applications (by applicant) to USPTO; normalised using max for cohort; multiplied by PA_USPTO_fa; only granted patents	Own calculations based on OECD Patent Quality (USPTO) Indicators (variable: quality_index_6) and OECD Patent Quality (USPTO) Indicators Cohort (for normalisation)

Note: The description of all the variables from ETER that can be merged with KC-HEI using the crosswalk *eter-to-patstat* (see Appendix B) is available at https://eter-project.com/data-for-download-and-visualisations/data-definitions/ and in ETER Handbook (Lepori, 2023)

Table 2A. Summary statistics of key indicators of university patenting

	Mean	Std. Dev.	Min	Max
Patent applications IP5	1.543	2.809	.014	75.949
Patent applications IP5 (FA*)	.982	1.818	.002	57.122
Patent applications EPO (FA*)	.586	1.007	0	23.939
Patent applications USPTO (FA*)	.339	.851	0	35.669
Patent applications IP5, granted	1.154	1.74	.011	36.742
Patent applications IP5, granted (FA*)	.695	1.066	.002	27.067
Patent applications EPO, granted (FA*)	.365	.597	0	12.071
Patent applications USPTO, granted (FA*)	.296	.598	0	20.67
AI patents IP	.019	.206	0	18.599
Forward citations EPO patents	.008	.031	0	1.936
Forward citations USPTO patents	.003	.024	0	2.844
Breakthrough EPO patents	0	.019	0	2.246
Breakthrough USPTO patents	.002	.045	0	5.914
EPO patents Quality (4components)	.048	.086	0	.984
USPTO patents Quality (4 components)	.036	.07	0	.783

Note: FA - fractional apportionment. IP5, EPO, USPTO stand for patent offices: Breakthrough patents identified as top 1% cited patents. Indicators of patent quality based on OECD Patent Quality database (Squicciarini et al., 2013). All indicators described in Table 1A

Source: own calculations using KC-HEI (Parteka et al., 2024)

Table 3A. List of patent tech fields (cpc3)

	cpc3 code	description
Human necessities	A01	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING
	A21	BAKING; EDIBLE DOUGHS
	A22	BUTCHERING; MEAT TREATMENT; PROCESSING POULTRY OR FISH
	A23	FOODS OR FOODSTUFFS; TREATMENT THEREOF, NOT COVERED BY OTHER CLASSES
	A24	TOBACCO; CIGARS; CIGARETTES; SIMULATED SMOKING DEVICES; SMOKERS' REQUISITES
	A41	WEARING APPAREL
	A42	HEADWEAR
	A43	FOOTWEAR
	A44	HABERDASHERY; JEWELLERY
	A45	HAND OR TRAVELLING ARTICLES
	A46	BRUSHWARE
	A47	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES; COFFEE MILLS; SPICE MILLS; SUCTION CLEANERS IN GENERAL
	A61	MEDICAL OR VETERINARY SCIENCE; HYGIENE
	A62	LIFE-SAVING; FIRE-FIGHTING
	A63	SPORTS; GAMES; AMUSEMENTS
	A99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Performing operations; transporting	B01	PHYSICAL OR CHEMICAL PROCESSES OR APPARATUS IN GENERAL
	B02	CRUSHING, PULVERISING, OR DISINTEGRATING; PREPARATORY TREATMENT OF GRAIN FOR MILLING
	B03	SEPARATION OF SOLID MATERIALS USING LIQUIDS OR USING PNEUMATIC TABLES OR JIGS; MAGNETIC OR ELECTROSTATIC SEPARATION OF SOLID MATERIALS FROM SOLID MATERIALS OR FLUIDS; SEPARATION BY HIGH-VOLTAGE ELECTRIC FIELDS
	B04	CENTRIFUGAL APPARATUS OR MACHINES FOR CARRYING-OUT PHYSICAL OR CHEMICAL PROCESSES
	B05	SPRAYING OR ATOMISING IN GENERAL; APPLYING FLUENT MATERIALS TO SURFACES, IN GENERAL
	B06	GENERATING OR TRANSMITTING MECHANICAL VIBRATIONS IN GENERAL
	B07	SEPARATING SOLIDS FROM SOLIDS; SORTING
	B08	CLEANING
	B09	DISPOSAL OF SOLID WASTE; RECLAMATION OF CONTAMINATED SOIL
	B21	MECHANICAL METAL-WORKING WITHOUT ESSENTIALLY REMOVING MATERIAL; PUNCHING METAL
	B22	CASTING; POWDER METALLURGY
	B23	MACHINE TOOLS; METAL-WORKING NOT OTHERWISE PROVIDED FOR
	B24	GRINDING; POLISHING
	B25	HAND TOOLS; PORTABLE POWER-DRIVEN TOOLS; MANIPULATORS
	B26	HAND CUTTING TOOLS; CUTTING; SEVERING
	B27	WORKING OR PRESERVING WOOD OR SIMILAR MATERIAL; NAILING OR STAPLING MACHINES IN GENERAL
	B28	WORKING CEMENT, CLAY, OR STONE
	B29	WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL
	B30	PRESSES
	B31	MAKING ARTICLES OF PAPER, CARDBOARD OR MATERIAL WORKED IN A MANNER ANALOGOUS TO PAPER; WORKING PAPER, CARDBOARD OR MATERIAL WORKED IN A MANNER ANALOGOUS TO PAPER
	B32	LAYERED PRODUCTS
	B33	ADDITIVE MANUFACTURING TECHNOLOGY
	B41	PRINTING; LINING MACHINES; TYPEWRITERS; STAMPS
	B42	BOOKBINDING; ALBUMS; FILES; SPECIAL PRINTED MATTER

	B43	WRITING OR DRAWING IMPLEMENTS; BUREAU ACCESSORIES
	B44	DECORATIVE ARTS
	B60	VEHICLES IN GENERAL
	B61	RAILWAYS
	B62	LAND VEHICLES FOR TRAVELLING OTHERWISE THAN ON RAILS
	B63	SHIPS OR OTHER WATERBORNE VESSELS; RELATED EQUIPMENT
	B64	AIRCRAFT; AVIATION; COSMONAUTICS
	B65	CONVEYING; PACKING; STORING; HANDLING THIN OR FILAMENTARY MATERIAL
	B66	HOISTING; LIFTING; HAULING
	B67	OPENING, CLOSING {OR CLEANING} BOTTLES, JARS OR SIMILAR CONTAINERS; LIQUID HANDLING
	B68	SADDLERY; UPHOLSTERY
	B81	MICROSTRUCTURAL TECHNOLOGY
	B82	NANOTECHNOLOGY
	B99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Chemistry; metallurgy	C01	INORGANIC CHEMISTRY
	C02	TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE
	C03	GLASS; MINERAL OR SLAG WOOL
	C04	CEMENTS; CONCRETE; ARTIFICIAL STONE; CERAMICS; REFRACTORIES
	C05	FERTILISERS; MANUFACTURE THEREOF
	C06	EXPLOSIVES; MATCHES
	C07	ORGANIC CHEMISTRY
	C08	ORGANIC MACROMOLECULAR COMPOUNDS; THEIR PREPARATION OR CHEMICAL WORKING-UP; COMPOSITIONS BASED THEREON
	C09	DYES; PAINTS; POLISHES; NATURAL RESINS; ADHESIVES; COMPOSITIONS NOT OTHERWISE PROVIDED FOR; APPLICATIONS OF MATERIALS NOT OTHERWISE PROVIDED FOR
	C10	PETROLEUM, GAS OR COKE INDUSTRIES; TECHNICAL GASES CONTAINING CARBON MONOXIDE; FUELS; LUBRICANTS; PEAT
	C11	ANIMAL OR VEGETABLE OILS, FATS, FATTY SUBSTANCES OR WAXES; FATTY ACIDS THEREFROM; DETERGENTS; CANDLES
	C12	BIOCHEMISTRY; BEER; SPIRITS; WINE; VINEGAR; MICROBIOLOGY; ENZYMOLOGY; MUTATION OR GENETIC ENGINEERING
	C13	SUGAR INDUSTRY
	C14	SKINS; HIDES; PELTS; LEATHER
	C21	METALLURGY OF IRON
	C22	METALLURGY; FERROUS OR NON-FERROUS ALLOYS; TREATMENT OF ALLOYS OR NON-FERROUS METALS
	C23	COATING METALLIC MATERIAL; COATING MATERIAL WITH METALLIC MATERIAL; CHEMICAL SURFACE TREATMENT; DIFFUSION TREATMENT OF METALLIC MATERIAL; COATING BY VACUUM EVAPORATION, BY SPUTTERING, BY ION IMPLANTATION OR BY CHEMICAL VAPOUR DEPOSITION, IN GENERAL; INHIBITING CORROSION OF METALLIC MATERIAL OR INCRUSTATION IN GENERAL
	C25	ELECTROLYTIC OR ELECTROPHORETIC PROCESSES; APPARATUS THEREFOR
	C30	CRYSTAL GROWTH
	C40	COMBINATORIAL TECHNOLOGY
	C99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Textiles; paper	D01	NATURAL OR MAN-MADE THREADS OR FIBRES; SPINNING
	D02	YARNS; MECHANICAL FINISHING OF YARNS OR ROPES; WARPING OR BEAMING
	D03	WEAVING
	D04	BRAIDING; LACE-MAKING; KNITTING; TRIMMINGS; NON-WOVEN FABRICS
	D05	SEWING; EMBROIDERING; TUFTING
	D06	TREATMENT OF TEXTILES OR THE LIKE; LAUNDERING; FLEXIBLE MATERIALS NOT OTHERWISE PROVIDED FOR
	D07	ROPES; CABLES OTHER THAN ELECTRIC
	D10	INDEXING SCHEME ASSOCIATED WITH SUBLASSES OF SECTION D,

		RELATING TO TEXTILES
	D21	PAPER-MAKING; PRODUCTION OF CELLULOSE
	D99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Fixed constructions	E01	CONSTRUCTION OF ROADS, RAILWAYS, OR BRIDGES
	E02	HYDRAULIC ENGINEERING; FOUNDATIONS; SOIL SHIFTING
	E03	WATER SUPPLY; SEWERAGE
	E04	BUILDING
	E05	LOCKS; KEYS; WINDOW OR DOOR FITTINGS; SAFES
	E06	DOORS, WINDOWS, SHUTTERS, OR ROLLER BLINDS IN GENERAL; LADDERS
	E21	EARTH OR ROCK DRILLING; MINING
	E99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	F01	MACHINES OR ENGINES IN GENERAL; ENGINE PLANTS IN GENERAL; STEAM ENGINES
	F02	COMBUSTION ENGINES; HOT-GAS OR COMBUSTION-PRODUCT ENGINE PLANTS
	F03	MACHINES OR ENGINES FOR LIQUIDS; WIND, SPRING, OR WEIGHT MOTORS; PRODUCING MECHANICAL POWER OR A REACTIVE PROPULSIVE THRUST, NOT OTHERWISE PROVIDED FOR
	F04	POSITIVE - DISPLACEMENT MACHINES FOR LIQUIDS; PUMPS FOR LIQUIDS OR ELASTIC FLUIDS
	F05	INDEXING SCHEMES RELATING TO ENGINES OR PUMPS IN VARIOUS SUBCLASSES OF CLASSES F01-F04
	F15	FLUID-PRESSURE ACTUATORS; HYDRAULICS OR PNEUMATICS IN GENERAL
	F16	ENGINEERING ELEMENTS AND UNITS; GENERAL MEASURES FOR PRODUCING AND MAINTAINING EFFECTIVE FUNCTIONING OF MACHINES OR INSTALLATIONS; THERMAL INSULATION IN GENERAL
	F17	STORING OR DISTRIBUTING GASES OR LIQUIDS
	F21	LIGHTING
	F22	STEAM GENERATION
	F23	COMBUSTION APPARATUS; COMBUSTION PROCESSES
	F24	HEATING; RANGES; VENTILATING
	F25	REFRIGERATION OR COOLING; COMBINED HEATING AND REFRIGERATION SYSTEMS; HEAT PUMP SYSTEMS; MANUFACTURE OR STORAGE OF ICE; LIQUEFACTION SOLIDIFICATION OF GASES
	F26	DRYING
	F27	FURNACES; KILNS; OVENS; RETORTS
	F28	HEAT EXCHANGE IN GENERAL
	F41	WEAPONS
	F42	AMMUNITION; BLASTING
	F99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Physics	G01	MEASURING; TESTING
	G02	OPTICS
	G03	PHOTOGRAPHY; CINEMATOGRAPHY; ANALOGOUS TECHNIQUES USING WAVES OTHER THAN OPTICAL WAVES; ELECTROGRAPHY; HOLOGRAPHY
	G04	HOROLOGY
	G05	CONTROLLING; REGULATING
	G06	COMPUTING; CALCULATING OR COUNTING
	G07	CHECKING-DEVICES
	G08	SIGNALLING
	G09	EDUCATION; CRYPTOGRAPHY; DISPLAY; ADVERTISING; SEALS
	G10	MUSICAL INSTRUMENTS; ACOUSTICS
	G11	INFORMATION STORAGE
	G12	INSTRUMENT DETAILS
	G16	INFORMATION AND COMMUNICATION TECHNOLOGY [ICT] SPECIALLY ADAPTED FOR SPECIFIC APPLICATION FIELDS
	G21	NUCLEAR PHYSICS; NUCLEAR ENGINEERING
	G99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Electricity	H01	ELECTRIC ELEMENTS

	H02	GENERATION; CONVERSION OR DISTRIBUTION OF ELECTRIC POWER
	H03	ELECTRONIC CIRCUITRY
	H04	ELECTRIC COMMUNICATION TECHNIQUE
	H05	ELECTRIC TECHNIQUES NOT OTHERWISE PROVIDED FOR
	H10*	SEMICONDUCTOR DEVICES; ELECTRIC SOLID-STATE DEVICES NOT OTHERWISE PROVIDED FOR
	H99*	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
Emerging Cross-Sectional Technologies	Y02	TECHNOLOGIES OR APPLICATIONS FOR MITIGATION OR ADAPTATION AGAINST CLIMATE CHANGE
	Y04	INFORMATION OR COMMUNICATION TECHNOLOGIES HAVING AN IMPACT ON OTHER TECHNOLOGY AREAS
	Y10	TECHNICAL SUBJECTS COVERED BY FORMER USPC

Note: CPC3 codes marked with * are not present in KC-HEI

Table 4A. List of 866 universities (HEIs) in Sample 1 (KC-HEI, 1980-2019)
In separate Excel sheet: Sample1.xlsx

Table 5A. List of universities (HEIs) in Sample 2 (2011-2019)
In separate Excel sheet: Sample2.xlsx

Appendix B

Description of the code constructing a crosswalk between the identifiers of HEIs (Higher Education Institutions) in PATSTAT (psn_id) and in ETER (eter_id)

Code (script) file: eter-to-patstat.ipynb (annotated version: eter-to-patstat.ipynb)[49]

Crosswalk file: *crosswalk_PSNid_to_ETERid.csv*

Abstract

The appendix describes the procedure leading to the creation of a crosswalk between the identifiers of entities (here: Higher Education Institutions, HEIs) in PATSTAT (variable: psn_id) and in ETER (variable: eter_id). The crosswalk enables the match between patent-level records in PATSTAT with the characteristics of universities (as patent applicants) provided by ETER or any other database using eter_id as university identifier. The correspondence is provided for **1068** HEIs from **31** European countries.

I. Data sources

The crosswalk is created using the data from:

1) PATSTAT Global – 2022 Autumn edition
   https://www.epo.org/searching-for-patents/business/patstat.html
2) ETER - European Tertiary Education Register (https://www.eter-project.com/);
   date of download: Sept 18, 2023

II. Crosswalk creation procedure - general description

The code first imports the necessary libraries and loads the dataset from PATSAT 'Tls206_person.csv' and ETER database. They are processed and filtered to include only the data meeting specific criteria (e.g. patent applications originating from universities from selected European countries) and selected variables (columns) needed for further analysis. The next part of the code calculates similarity between the university names in PATSTAT and in ETER. For each row in the Tls206_person dataset, it searches for similar names in the ETER dataset using a string similarity metric. The corresponding eter_id and name are stored as the most likely match. After calculating the similarities and choosing the highest one, the results are saved in a CSV file. The full script consists of the following steps (tasks) - described below.

1. Step 1. Libraries and Data

[49] AI-optimised version of the code is stored as eter-to-patstat_AI.ipynb (annotated version: eter-to-patstat_AI.pdf).

2. Step 2. Transformation of ETER database
3. Step 3. Transformation of dataset 'Tls206_person'
4. Step 4. Checking for similarity
5. Step 5. Uniforming the records
6. Step 6. Manual check
7. Step 7. Final crosswalk creation

III. Steps description

III.1 Description of Step 1. Libraries and Data

The code imports several libraries and files required to function properly. The *PANDAS* package serves as a Python library for data manipulation and analysis. It offers a variety of data structures and functions tailored for efficient handling of structured data like tables and time series. The *NUMPY* package is used for scientific computing in Python. It facilitates operations on arrays, matrices, and mathematical functions designed to work efficiently with these data structures. The *MATH* package provides a collection of mathematical functions, encompassing basic arithmetic operations, trigonometric functions, logarithmic functions, and more. The *RE* package furnishes support for regular expressions, powerful tools for pattern matching and manipulating strings. The *DIFFLIB* package offers classes and functions tailored for comparing sequences, particularly for identifying differences between them. It is commonly employed for tasks such as computing dissimilarities between strings or sequences of lines in files. The *UNIDECODE* package, a Python library, specialises in converting Unicode data (typically text) into ASCII equivalents. This conversion process is frequently employed for standardising text data, facilitating tasks like string comparison or indexing.

III.2 Description of step 2: Transformation of ETER database

In this step, the code executes a series of operations on a DataFrame named 'eter'. Initially, it selectively extracts specific columns from the DataFrame, retaining only those relevant to the subsequent analysis. Then, a renaming procedure is employed to assign new labels to the selected columns, as defined within a dictionary structure. Following this, the code utilises the 'make_simple_name' function twice to generate additional columns, denoted as 'simple_name_1' and 'simple_name_2'. These new columns are derived from existing data in the 'name' and 'eng_name' columns, respectively, through a process of simplification or transformation. Moreover, the code undergoes an iterative process over each row of the

DataFrame to update country codes. Within this iteration, if the value within the 'ctry_code' column corresponds to 'UK', the code appends 'GB' to a designated result list. Conversely, for non-'UK' entries, the existing 'ctry_code' values are appended to the result list. Subsequently, the 'ctry_code' column in the DataFrame is revised to reflect the updated values stored within the result list, completing the data manipulation process.

III.3 Description of step 3: Transformation of dataset 'Tls206_person'

In this step, firstly, the code filters a DataFrame named 'person' to retain only rows where the value in the 'psn_sector' column is 'UNIVERSITY'. Then, it creates a list named 'countries_list' containing country codes. Next, it links up the elements of 'countries_list' into a single string separated by '|' symbols - for the purpose of further manipulations. Finally, it filters the 'person' DataFrame again, keeping only the rows where the 'person_ctry_code' column contains any of the country codes from the 'countries_list'.

Then, in the second substep, the code encompasses the definition of two essential functions, namely 'extract_postcode' and 'extract_eng_postcode'. Both the extract_postcode and extract_eng_postcode functions are used to extract postal codes from an address string. They are designed to handle different formats of postal codes based on the country's postal code rules. The extract_postcode function is used to extract postal codes from an address string for non-UK countries. The address parameter is the input address string from which the postal code needs to be extracted. The function uses a regular expression (postcode_regex) to find and extract postal code patterns from the address. The regular expression included in the function is designed to match postal codes that consist of 4 to 7 alphanumeric characters, optionally followed by 3 more alphanumeric characters separated by whitespace. The function returns the extracted postal code as a string and applies the regular expression to the address string. If it finds any matches, it returns the last match (assuming the last match is the most relevant), and removes any whitespace characters from the extracted postal code. If no matching postal code is found in the address, the function returns an empty string. The extract_eng_postcode function was designed accordingly, creating different expressions to extract specific english postcodes.

Then, all the names of the Universities (HEIs) from the PATSTAT's Tls206_person dataset are being simplified to increase their similarity and avoid the issue of typos in both datasets. Simplifying algorithm takes three arguments: (i) df: The DataFrame in which the name column exists; (ii) column_name: The name of the column that contains the names to be simplified; (iii) simple_name: The name of the new column where the simplified names will be

stored. The purpose of the function is to clean and simplify the names in the specified column by applying the following steps:

(1) Unicoding the name: It converts any accented characters or special characters to their closest ASCII representation using the unidecode library. For example, it would convert "é" to "e" and "ø" to "o".

(2) Removing non-letter characters: It uses a regular expression to remove any characters that are not letters (alphabets) from the name. This step removes punctuation, numbers, and any other non-letter characters.

(3) Converting to lowercase: It converts the entire name to lowercase letters. This step ensures that name comparisons are case-insensitive and more effective.

The function then creates a new column in the DataFrame with the simplified names and returns the modified DataFrame. By performing these steps, the make_simple_name function standardises the names in the DataFrame column, making them more suitable for comparison or analysis tasks. For instance, it can be used to simplify names in both the ETER and PATSTAT Data DataFrames before applying string similarity metrics (using the String_similarity function) to match individuals' records to institutions' records accurately. This helps in data integration and ensures more accurate matching results when dealing with names that may have variations due to the different formats, special characters, or capitalizations.

III.4 Description of step 4. Checking for similarity

In this step, the code matches HEIS/universities from PATSTAT's dataset (Tls206_person) with corresponding entities in the ETER database, based on similarities in their names. The procedure in the code follows the following steps:

(1) initialization - the script initialises an empty list called results. This list will be used to store the matching results between individuals from the person dataset and entries in the ETER dataset.

(2) extracting information about each record - the script extracts different attributes of each entity in the database retrieved from the Tls206_person database, such as: country code (ctry_code), NUTS code (nuts3), postal code (postcode), and various identifiers and names.

(3) data filtering - the script filters ETER records based on the country code (ctry_code), NUTS code (nuts3), and postal code (postcode) determined by the current row of the person record. This procedure is performed using a Boolean index to select rows from the ETER dataset that match the specified criteria.

(4) duplicate removal - the script removes any duplicate rows from the filtered ETER dataset to ensure that each entity is considered only once in the matching process. This is done using the drop_duplicates() method.

(5) matching process - The script loops through each row in the filtered ETER dataset. For each row, the name associated with the person in the Tls206_person dataset from PATSTAT is compared to the name in the ETER dataset to determine the degree of similarity. The script assigns Winner_Level, Winner_eter_id, and Winner_name based on the highest similarity found between the names. String similarity comparison is then used to compare different name combinations in the two datasets. If a match is found, the winner_level, winner_eter_id, and Winner_name values are updated accordingly. The actual matching algorithm is based on comparing string similarity using the Levenstein distance.

(6) storing results - for each person in the people dataset, the script creates a dictionary called partial_result that contains the relevant identifiers (e.g. psn_id, han_id, person_id), name, associated winner_eter_id, and similarity (winner_level). This dictionary represents the matching results for the current person. Then the "partial_result" dictionary will be appended to the results list and all matching results collected.

(7) final dataset creation - once all individuals in the Tls206_person dataset have been processed and their matching results stored in the results list, the script creates a pandas DataFrame (df) from the list of results. This DataFrame will contain the final matching results, making it easier to further analyse and manipulate the data.

III.5 Description of step 5. Uniforming the records

In this step, the code aims to uniform the results, as many of the entities in PATSTAT's Tls206_person dataset are in fact the same entities, as this dataset is not perfectly uniformed. The code iterates through the index list (index_list). For each index, it extracts a subset of the data from the DataFrame new_df and sets default values for the variables eter_id, eter_name, patstat, and level. If the extracted data is a Pandas series, it assigns values based on specific columns. If it is a DataFrame it will look for the row with the highest similarity_ratio value. If the value is less than 1.0, the row with the most common eter_id is found. Then it creates dictionary data with related values and creates a DataFrame result_df. Finally, result_df is added to the existing DataFrame uni_df.

III.6 **Description of step 6. Manual check**

If similarity >=0.85 it is assumed that the names correspond to the same entity. If similarity is >0.45 and <0.85, we rely on the expert knowledge, performing additional manual and web checks, assigning a match if PATSTAT name (id) and ETER name (id) describe the same HEI.

The remaining 2514 PATSTAT names (with similarity ratio below 0.45; or higher but identified in the previous step as incorrectly matched by the algorithm) went through an extensive manual matching with the aid of ETER database and web search. Around 64% of these records were successfully matched thanks to this procedure (with 911 records remaining as unmatched).

Some cases failed to be matched automatically because the entity names contained additional information (such as the name of the faculty/department/institute/other) or the name was very "messy" (typos; university name written neither in original language nor in English but in some other language). Another case of the algorithm failure, easily solved manually, was when HEI operates under various names or abbreviations or if the PATSTAT entity A is an entity dependent from/affiliated to a given ETER entity B. Web search also helped to match PATSTAT to ETER in case of HEIs changing name in time. At this stage we did not take into account the time dimension of the data. ETER data starts from 2011 which implies that, for example, if A changes name to B in 2005, PATSTAT entity A will be assigned to ETER entity B (since only B exists in ETER). If one uses PATSTAT data merged with ETER (available for 2011-2020) this matched records will be dropped.

Other similar cases of entities existing in PATSTAT but not in ETER were also matched using web search. This could be the case of PATSTAT entity A being taken over by ETER entity B at some point of time or PATSTAT entity A merged with some other entities to become ETER entity B. The opposite case, i.e. PATSTAT entity A splitting into many entities existing in ETER could not be matched.

Additionally, university hospitals affiliated to one HEI were assigned to this HEI. Similarly, technology transfer offices (TTOs) with clear affiliation/owned by a HEI were assigned to this HEI.

Entities affiliated to more than one ETER records (like many universities' consortia) were left unmatched due to the difficulty of assessing the exact shares. Among the unmatched records we have found also some units incorrectly matched in PATSTAT as HEIs, for instance private companies without any connection to HEI, or private persons with unidentified affiliation (if any) to a HEI.

Outcome: file "*full_PSNid_to_ETERid.csv*" .

III.7 Task 7. Final crosswalk creation

In this final step, the script performs a series of operations on a DataFrame named uni_df. First, the code imports the data from the CSV file "*full_PSNid_to_ETERid.csv*" (the file after manual check), using "|" as a separator. It then ensures that the eter_id column in uni_df is converted to string type for consistency. The code then filters uni_df to retain only rows where the value in the matches_names column is equal to 1, indicating a successful match. After this filtering, only the "psn_id" and "eter_id" columns are selectively extracted from the DataFrame. To ensure data integrity, all duplicate rows are removed based on the combination of psn_id and eter_id, thereby eliminating redundancy. Then reset the DataFrame's index to maintain order. Finally, the code uses "|" to export the processed DataFrame to a new CSV file named "crosswalk_PSNid_to_ETERid.csv". As a delimiter, thus encapsulating the delicate data extraction and manipulation process.

Outcome: file *crosswalk_PSNid_to_ETERid.csv*: a crosswalk between HEIs identifiers in PATSTAT (psn_id) and in ETER (eter_id) to be used to merge patent records from PATSTAT with HEIs characteristics from ETER or other databases using eter-id identifier.

Original citation:
Parteka P., Płatkowski P., Szymczak S., Wolszczak-Derlacz J. (2024). A dataset on knowledge creation and patenting by European Higher Education Institutions (KC-HEI). GUT FME Working Papers Series A, No 2/2024(73). Gdansk (Poland): Gdansk University of Technology, Faculty of Management and Economics.

All GUT Working Papers are downloadable at:

http://zie.pg.edu.pl/working-papers

GUT Working Papers are listed in Repec/Ideas
https://ideas.repec.org/s/gdk/wpaper.html

Gdańsk University of Technology, Faculty of Management and Economics
Narutowicza 11/12, (premises at ul. Traugutta 79)
80-233 Gdańsk, phone: 58 347-18-99 Fax 58 347-18-61
www.zie.pg.edu.pl

GDAŃSK UNIVERSITY OF TECHNOLOGY

FACULTY OF MANAGEMENT AND ECONOMICS